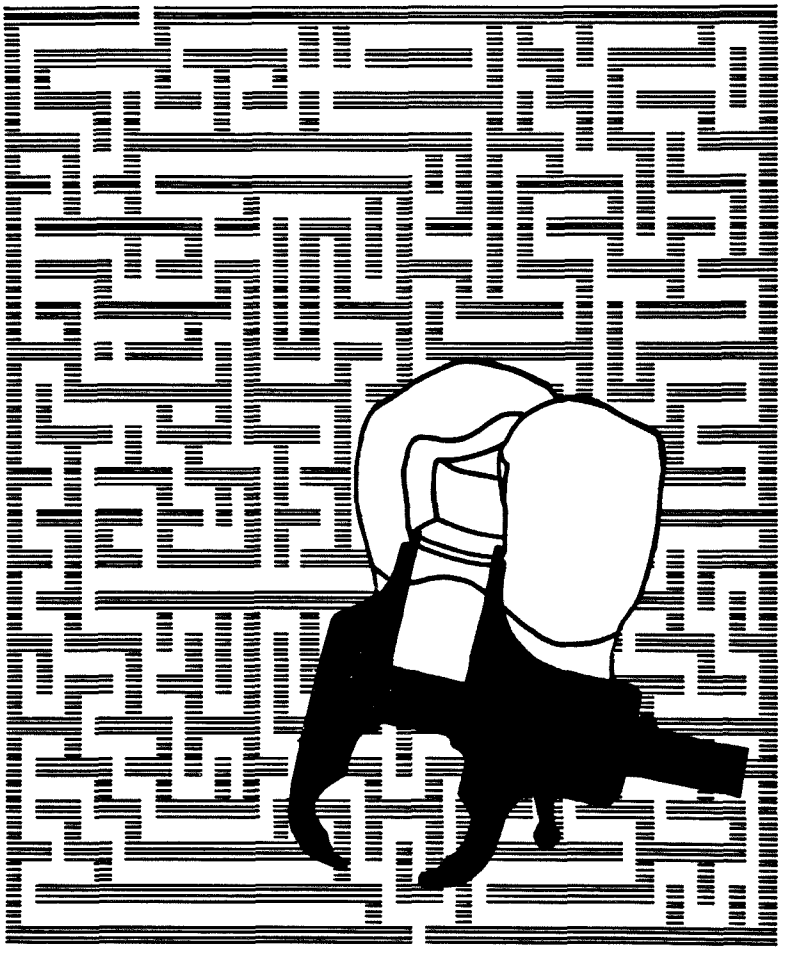
EVALUATIE IN HET TANDHEELKUNDIG ONDERWIJS

beoordelen van practicumwerkstukken en
meten van probleemoplosvaardigheid



GERARD STRAETMANS

EVALUATIE IN HET TANDHEELKUNDIG ONDERWIJS

beoordelen van practicumwerkstukken en
meten van probleemoplosvaardigheid

EVALUATION IN DENTAL EDUCATION

assessment of preclinical performance
and problem solving ability

Promotores: Prof. Dr. A.J.M. Plasschaert
Prof. Dr. D.W. Vaags

Uit het Instituut Conserverende Tandheelkunde voor
Volwassenen van de Katholieke Universiteit te Nijmegen.
Hoofd: Prof. Dr. A.J.M. Plasschaert

CIP-GEGEVENS KONINKLIJKE BIBLIOTHEEK, DEN HAAG

Straetmans, Gerardus Josephus Johannes Maria

Evaluatie in het tandheelkundig onderwijs: beoordelen van practicumwerkstukken en meten van probleemoplosvaardigheid / Gerardus Josephus Johannes Maria Straetmans. - [S.l.: s.n.]. - Ill.
Proefschrift Nijmegen. - Met lit. opg.
ISBN 90-9000894-2
SISO 609.3 UDC 378.012:616.314
Trefw.: tandheelkundig onderwijs.

EVALUATIE IN HET TANDHEELKUNDIG ONDERWIJS

beoordelen van practicumwerkstukken en
meten van probleemoplosvaardigheid

EVALUATION IN DENTAL EDUCATION

assessment of preclinical performance
and problem solving ability

(with a summary in English)

PROEFSCHRIFT

ter verkrijging van de graad van doctor
in de geneeskunde
aan de Katholieke Universiteit te Nijmegen
op gezag van de Rector Magnificus
Prof. Dr. J. H. G. I. Giesbers
volgens besluit van het College van Dekanen
in het openbaar te verdedigen
op vrijdag 10 mei 1985
des namiddags om 2.00 uur precies

door

Gerardus Josephus Johannes Maria Straetmans

geboren te Nijmegen

krips repro meppel

Every child should have equal
opportunity to be treated unequally

T. Husén

Voor Kitty
Voor mijn moeder en ter na-
gedachtenis aan mijn vader

Aan de totstandkoming van dit proefschrift hebben vele personen op directe of indirecte wijze een bijdrage geleverd.
In de eerste plaats moeten in dit verband de collega's van het Instituut genoemd worden die, volgens het universiteitsreglement, hier niet met name genoemd mogen worden. Voor hun medewerking en belangstelling ben ik veel dank verschuldigd.

Veel studenten werkten mee aan het testen van de ontwikkelde evaluatie-instrumenten en hebben daarmee bijgedragen aan de kwaliteitsverbetering van het tandheelkundig onderwijs. Ik ben ze zeer erkentelijk voor hun kritische participatie.

Wim Vaags hielp mij om mijn onderwijskundige kennis te activeren en deze efficiënt aan te wenden in het tandheelkundig domein. Ik dank hem voor zijn deskundige, accurate en prettige wijze van begeleiden.

Rietje Walker-Mallens verdient lof voor de deskundige wijze waarop zij de samenvatting in dit proefschrift vertaalde in het Engels.

Harry Reckers, Jos van der Kamp en Henk Bongaarts droegen zorg voor de figuren in dit proefschrift.

De persoon die hier zeker genoemd moet worden is Kitty, mijn voornaamste bron van inspiratie. De vrije tijd van de afgelopen jaren, geïnvesteerd in dit proefschrift, was geen persoonlijk bezit. De schuld die ik heb opgebouwd doordat ik er toch als zodanig over kon beschikken, kan slechts ten dele worden ingelost met de belofte dat het de komende jaren anders zal zijn.

INHOUD

ALGEMENE INLEIDING

Het centrale thema in deze dissertatie betreft de evaluatie in het tandheelkundig onderwijs. Volgens De Corte et al. (1976) is het doel van evaluatie bij te dragen tot verbetering of optimalisering van het didactisch proces. Deze bijdrage ligt dan vooral in het verzamelen van informatie op basis waarvan beslissingen genomen kunnen worden over het onderwijs en/of de studenten. In dit proefschrift gaat het primair over het optimaliseren van te nemen beslissingen over studenten. Dergelijke beslissingen worden genomen met het oogmerk van selectie of diagnose. Selectie is er op gericht om de kwaliteit van het onderwijs te waarborgen, terwijl diagnose meer betrekking heeft op de efficiëntie van dat onderwijs. Ongeacht wat het doel is, moet staat gemaakt kunnen worden op de verzamelde informatie. Daarvoor is nodig dat instrumenten waarmee de informatie verzameld wordt valide zijn (dat wil zeggen dat de juiste informatie verkregen wordt) en op de voorgeschreven wijze gebruikt worden. Te vaak komt het voor dat niet of slechts gedeeltelijk aan deze eisen voldaan wordt. Ook aan de Nijmeegse Subfaculteit Tandheelkunde wordt regelmatig geconstateerd dat de kwaliteit van de evaluatie-instrumenten te wensen overlaat. Over recente pogingen om valide instrumenten te construeren voor het vaststellen van psychomotorische vaardigheden en probleemoplossend vermogen handelt dit proefschrift. De onderzoeks- en ontwikkelingswerkzaamheden zijn uitgevoerd in twee onderwijsstimuleringsprojecten die verworven waren door het Instituut Conserverende Tandheelkunde voor Volwassenen van de Katholieke Universiteit te Nijmegen. De verrichte werkzaamheden in deze projecten zijn beschreven in de twee delen van dit proefschrift, welke onafhankelijk van elkaar gelezen kunnen worden.

In deel I ligt de nadruk op het vaststellen van psychomotorische vaardigheden. Getracht is om de beoordelingskwaliteit van preklinische practicumwerkstukken te verbeteren door de constructie van een nieuw beoordelingsinstrument alsmede door de ontwikkeling van een trainingsprogramma voor beoordelaars.

In deel II gaat de aandacht uit naar het vaststellen van probleemoplosvaardigheid. Ten einde meer betrouwbare beslissingen te kunnen nemen over de toelating van studenten tot klinische patiëntbehandeling, zijn nieuwe evaluatie-instrumenten ontwikkeld en getest. De ontwikkelde instrumenten zijn vormen van "papieren simulatie" en beschikken mede daardoor over eigenschappen die ze tevens geschikt maken als leermiddel.

DEEL I

BEOORDELEN VAN PRACTICUMWERKSTUKKEN

I MOTORISCHE VAARDIGHEDEN IN HET TANDHEELKUNDIG ONDERWIJS

1.1 Inleiding

In de opleiding voor tandarts neemt het aanleren van motorische vaardigheden een belangrijke plaats in. Dit komt tot uiting in de hoeveelheid tijd, die in het curriculum wordt uitgetrokken voor het motorische onderwijs. Aan de Nijmeegse subfaculteit Tandheelkunde wordt ongeveer 70 procent van het totale contactonderwijs besteed aan onderwijs in motorische vaardigheden (Otto, 1981). Deze vaardigheden worden zowel in de klinische als in de preklinische fase aangeleerd. In de eerstgenoemde onderwijssituatie worden patiënten behandeld, in laatstgenoemde worden studenten op die behandeling voorbereid. Deze voorbereiding bestaat uit het oefenen van allerlei technische vaardigheden in een gesimuleerde situatie.
Het bestaan van preklinisch motorisch onderwijs is gebaseerd op de veronderstelling dat dit onderwijs dié technische vaardigheden ontwikkelt, die essentiëel zijn voor het behandelen van patiënten. Silvestri et al.(1979) onderzochten deze assumptie en kwamen, op basis van het verschil in resultaten op een vóór- en natest, tot de conclusie dat de meeste studenten inderdaad hun technische vaardigheid hadden vergroot als gevolg van het preklinisch onderwijs. In een studie uitgevoerd door Massler en Evans (1977) komt echter naar voren, dat aan de waarde van preklinische vaardigheid als voorspeller van klinische vaardigheid, getwijfeld moet worden. Als oorzaak noemen de auteurs de wijze van beoordelen en met name het ongestandaardiseerde karakter daarvan.
In het reeds aangehaalde artikel van Silvestri wordt eveneens de vraag gesteld naar de validiteit van het beoordelingssysteem. De auteurs twijfelen aan de waarde van relatieve (ten opzichte van medestudenten) prestatie-oordelen, omdat deze de werkelijk toegenomen vaardigheid niet zouden reflecteren: ".....rank ordered grading alone is not sufficient in either identifying students who have improved their technical skill or identifying students who have reached their maximum technical skill potential. Another dimension of evaluation must accompany the routine grade to inform the student of technical skill development."
In een later verschenen artikel (Cohen et al.,1980) laten dezelfde auteurs zien dat het beoordelen van werkstukken aan de hand van een evaluatieformulier een positieve invloed heeft op de verwerving van motorische vaardigheden in het preklinisch onderwijs. Het ingevulde formulier gaf informatie over sterke en zwakke punten van de preparatie door middel van trefwoorden, lijnen en illustraties. Een enquête onder de betrokken studenten wees uit, dat de meesten van mening waren, dat het ingevulde formulier van groot nut was voor het verstrekken van gerichte terugkoppeling.
Dat terugkoppeling een erg belangrijk doel is van beoordelen wordt door zeer veel onderzoekers van het tandheelkundig onderwijs onderkend (Hinkelman en Long, 1973; Mackenzie, 1973; Salvendy et al., 1976; Schiff et al., 1975). Tegelijkertijd wordt echter ook

geconstateerd dat de terugkoppeling dikwijls van twijfelachtige kwaliteit is en daardoor het leren onvoldoende stimuleert.

In hoofdstuk II wordt nader onderzocht wat de voornaamste problemen zijn met betrekking tot het beoordelen van motorische vaardigheden in het tandheelkundig onderwijs en welke oplossingen daarvoor mogelijk zijn. In de resterende paragrafen van dit hoofdstuk wordt achtereenvolgens besproken wat onder vaardigheden verstaan wordt, hoe deze verworven worden en hoe vastgesteld kan worden of ze verworven zijn.

1.2 De verwerving van vaardigheden

1.2.1 Kenmerken van het verwervingsproces van vaardigheden

Wat is een vaardigheid? Het psychologisch gebruik van de term is tamelijk breed en omvat dié processen die vakkundige, snelle en accurate prestaties tot gevolg hebben. Deze definitie van vaardigheid is niet alleen van toepassing op taken die manuele activiteit inhouden, maar ook op dié welke interne manipulatie van symbolen vereisen, zoals bij het gebruik van taal (Fitts en Posner, 1967). Volgens Fitts et al.(1959) is het voornaamste kenmerk van een vaardigheid, de organisatie van gedragscomponenten in tijd en ruimte. Door leren en ervaring wordt een bepaalde, hoge organisatiegraad bereikt. Deze spatiële en temporele organisatie van het gedrag wordt gekenmerkt door:
- timing en anticipatie;
- constantie van respons en effect;
- het gebruik van teruggekoppelde informatie.

Timing en anticipatie zorgen ervoor dat de onderdelen waaruit een bepaalde vaardigheid bestaat in de juiste volgorde worden afgewerkt en dat ieder onderdeel op het juiste ogenblik volgt op het vorige. Vaardigheden die een vast en gedwongen tempo kennen (bijvoorbeeld het serveren van een tennisbal), kunnen alleen goed verlopen als het patroon van de stimuli geanticipeerd kan worden. De gerelateerdheid van timing en anticipatie blijkt uit de volgende omschrijving van het begrip timing: "de correcte anticipatie van het tijdstip van gebeurtenissen" (Mulder et al., 1976). Anticipatie zorgt ervoor dat ingewikkelde handelingen vloeiend uitgevoerd worden. Mulder et al. (1976) onderscheiden twee vormen van anticipatie: perceptieve en cognitieve anticipatie.
De perceptieve anticipatie kenmerkt zich doordat een persoon het verloop van de stimulus van tevoren kan zien aankomen. De automobilist* kan het verloop van de weg gewoonlijk enkele honderden meters vooruit waarnemen. De tandarts c.q. student tandheelkunde ziet tijdens het boren hoe ver de preparatie gevorderd is en waar het boren moet ophouden.

---

* Voor het gemak wordt in dit proefschrift steeds naar personen verwezen in de mannelijke vorm.

Cognitieve anticipatie doet zich voor als een persoon gebruik maakt van zijn verworven kennis omtrent de dynamische eigenschappen van de stimulus. De (a.s.) tandarts weet, op grond van een geheugenrepresentatie, hoe preparaties er meestal uitzien en waar de kans op eventuele complicaties het grootst is.

Onder constantie van responsie verstaan Mulder et al. (1976) de mogelijkheid om eenzelfde handelingsverloop te bewerkstelligen via geheel verschillende innervatiepatronen (beïnvloedingspatronen door het zenuwstelsel) van de spieren waarmee de handeling verricht wordt. Vermoeidheid als gevolg van herhaalde bewegingen kan daardoor worden uitgesteld.

Onder constantie van effect verstaan genoemde auteurs het verschijnsel dat ook bij variatie in de wijze waarop het gedrag wordt uitgevoerd, overeenkomstige resultaten worden verkregen. De tandarts (in opleiding) kan een klasse II preparatie maken in verschillende elementen. Uit het bestaan van deze constanties leiden de auteurs af, dat het leren van een motorische vaardigheid meer inhoudt dan het leren van specifieke bewegingen. Behalve het uitvoeringsschema voor de deelbewegingen waaruit de vaardigheid bestaat, zijn er een aantal regels die beschrijven welke volgordes en variaties in de volgorde zijn toegestaan. Bruner (1970) spreekt in dit verband van functioneel equivalente variaties die een vaardigheid produktief maken. Hij bedoelt daarmee dat met de regels van een ontwikkelde vaardigheid steeds nieuwe, nog niet eerder vertoonde, variaties op het bewegingsplan geproduceerd kunnen worden.

Terugkoppeling van informatie is van doorslaggevend belang voor de ontwikkeling van vaardigheden. Door terugkoppeling wordt het mogelijk het onderscheid tussen het beoogde doel van een handeling en het effect van die handeling te vergelijken.
Bij het leren van vaardigheden is behalve terugkoppeling van informatie over grootte en richting van de discrepantie (evaluatieve terugkoppeling), ook informatie nodig over de wijze waarop deze kan worden verkleind (informatieve terugkoppeling). Mulder et al.(1976) onderscheiden drie typen van terugkoppeling:

- intrinsieke terugkoppeling
  Deze vorm van terugkoppeling is een natuurlijk gevolg van de beweging zelf. De beweging leidt tot prikkeling van bepaalde perifere organen in de spieren en deze prikkeling wordt teruggevoerd naar het centrale zenuwstelsel, aldus informatie gevend over positie en snelheid van beweging van de ledematen. Tegelijkertijd wordt perifere informatie verkregen vanuit het auditieve en het visuele systeem: men hoort en ziet wat men doet.
- extrinsieke terugkoppeling
  Extrinsieke terugkoppeling kent verschillende vormen. In Behavioristische leertheorieën komt men het tegen als "beloning". Het proefdier krijgt een kleine hoeveelheid voedsel als het een gewenste gedraging vertoont.
  Een veel gebruikte vorm van extrinsieke terugkoppeling bij het aanleren van gecompliceerde taken is "kennis van de resultaten".

De lerende krijgt informatie over de geleverde prestatie, bijvoorbeeld in de vorm "goed" of "fout". De terugkoppeling kan elke gewenste vorm van nauwkeurigheid aannemen. Kennis van de resultaten is zeer belangrijk in de eerste fase van de verwerving van vaardigheden, wanneer het vaak lastig is om het handelingsverloop mondeling duidelijk te maken aan de student (Schmidt, 1975). De docent moet zich dan behelpen met aanduidingen als "goed" of "fout". Zonder kennis van de resultaten zal het leren in de beginfase erg moeizaam gaan.

- interne terugkoppeling
  Onder interne of centrale terugkoppeling verstaat men terugkoppeling binnen het zenuwstelsel. Deze maakt het mogelijk dat het gedrag gestuurd wordt en fouten gecorrigeerd worden nog vóórdat met de uitvoering van het gedrag begonnen is. Het bestaan van interne terugkoppeling is noodzakelijk voor de verklaring van het foutloos verlopen van zeer snelle bewegingen. Als een beweging zó snel verloopt dat informatie van de spieren en de zintuigen (intrinsieke terugkoppeling) niet tijdig in de hogere centra arriveert om de afloop van een beweging te kunnen beïnvloeden, dan moet aangenomen worden dat besturing van het bewegingspatroon centraal geschiedt (Mulder et al., 1976).

1.2.2 Fasen in het verwervingsproces

Fitts (1964) onderscheidde een aantal fasen in het verwervingsproces van motorische vaardigheden. In de cognitieve fase raakt de lerende bekend met de taak doordat hij de eisen leert kennen en leert om acht te slaan op belangrijke perceptuele gegevens. In deze fase wordt veel geleerd van gemaakte fouten. Terugkoppeling speelt dan ook een belangrijke rol in dit vroege stadium. Gedurende de fixatie fase worden de perceptueel-cognitieve processen langzamerhand gedomineerd door de toenemende perceptueel-motorische coördinatie. Informatie vanuit de spieren (kinesthetische terugkoppeling) is naast cognitie noodzakelijk voor het uitvoeren van de taak. De tweede fase is de langste omdat het bereiken van perceptueel-motorische coördinatie veel oefening vereist. Belangrijke vragen met betrekking tot het leren in deze fase zijn:

1. Is langdurige training te prefereren boven meerdere korte trainingen?
2. Moet de taak als geheel geoefend worden of in onderdelen?

Op de eerste vraag kan geantwoord worden, dat meerdere korte trainingen over het algemeen een beter resultaat opleveren. Het voordeel van gespreide oefening voor de efficiëntie van het leren werd al in 1914 aangetoond door Lyon (1914, geciteerd in Hayes, 1981). Het antwoord op vraag twee is afhankelijk van de aan te leren vaardigheid. Als de taak bestaat uit deelvaardigheden die synchroon moeten verlopen, dan moet veel geoefend worden op de onderlinge timing ervan. Het oefenen van de taak als geheel verdient dan de voorkeur. Zijn de deelvaardigheden onafhankelijk van elkaar dan is het beter om iedere component apart te oefenen (Mulder et al., 1976).

Door oefening groeit de integratie tussen perceptuele en kines-

thetische informatie en vangt de derde fase aan: de automatisatie. Geautomatiseerde vaardigheden worden uitgevoerd met een minimum aan fouten en aandacht. De handelingen voltrekken zich autonoom, dat wil zeggen niet meer onderworpen aan cognitieve controle en minder gevoelig voor interferentie van andere activiteiten. Interne terugkoppeling (zie par. 1.2.1) wordt steeds belangrijker. Kennis van het verwervingsproces van motorische vaardigheden is belangrijk als leidraad voor het inrichten van onderwijs. Aldus ingericht onderwijs kan echter geen garantie zijn voor een onbelemmerde voortgang van het leerproces. Het is daarom gewenst dat instructeurs in staat zijn op te sporen waar het leerproces stagneert. Diagnostische tests kunnen daarbij behulpzaam zijn. Voor het ontwikkelen van dergelijke tests is een analyse van het leerproces nuttig. Leren, of het nu motorische, cognitieve of affectieve vaardigheden betreft, kent drie componenten: de stimuli, de mediërende responsen (te weten dié responsen, die de stimulus en de waarneembare respons associëren) en de waarneembare responsen. Om de verwerving van motorische vaardigheden in het tandheelkundig onderwijs efficiënter te laten verlopen, beveelt Mackenzie (1973) aan, om die drie componenten onafhankelijk van elkaar te evalueren. In onderstaande beschrijving van de leercomponenten van motorische vaardigheden wordt Mackenzie gevolgd.

Stimuli

Vanuit hun werk krijgen studenten aanwijzingen, die hen ertoe moeten brengen om beslissingen te nemen met betrekking tot de volgende te ondernemen actie. Ze moeten in staat zijn om te bepalen wanneer de caviteitspreparatie het dentine heeft bereikt, wanneer al het aangetaste weefsel verwijderd is, wanneer er sprake is van onondersteund glazuur, enz. Geëxtraheerde elementen, waar een preparatie in is gemaakt, kunnen gebruikt worden om te bepalen of studenten de juiste aanwijzingen herkennen. De bekwaamheid van de student om significante aanwijzingen te onderscheiden, wordt op deze directe manier veel efficiënter vastgesteld dan door middel van het laten vervaardigen van een aantal preparaties.

Mediërende responsen

Het begrip van wat een acceptabele of onacceptabele caviteitspreparatie is fungeert als intermedium tussen de stimulus en de waarneembare respons. Studenten hebben een juist begrip van caviteitspreparaties als ze onderscheid kunnen maken tussen acceptabel en niet acceptabel en binnen de grenzen van het acceptabele de verscheidenheid van mogelijke caviteitspreparaties herkennen. Om dit onderscheid te kunnen maken moet men weten wat de relevante en irrelevante attributen zijn. Door testsituaties te creëren die met betrekking tot de relevante en irrelevante attributen zeer wijd uiteenlopen, kan snel nagegaan worden of een persoon het begrip duidelijk heeft.

Waarneembare responsen

Ondanks het feit dat een student significante aanwijzingen herkent en een goed begrip heeft van caviteitspreparaties, kan zijn

klinische prestatie toch te wensen overlaten. Het probleem kan dan liggen in de uitvoering. Het is mogelijk dat de student niet geleerd heeft om te gaan met de instrumenten, te haastig is of te onstuimig in het hanteren van het instrumentarium. Gestandaardiseerde testsituaties, waarin het gebruik van de instrumenten geanalyseerd wordt, kunnen de oorzaak helpen opsporen.

1.3 Het vaststellen van de verwerving van vaardigheden

1.3.1 Inleiding

Het uiteindelijke doel van alle onderwijs is dat de lerende iets kan of weet wat vóór het genoten onderwijs niet geweten of beheerst werd. Het uiteindelijke doel van het tandheelkundig onderwijs is klinische competentie. Wat daaronder verstaan wordt komt in par. 1.3.3 aan de orde. Tijdens en na afloop van het verwervingsproces moet beoordeeld worden in hoeverre de lerende "klinische competentie" bezit. Dit beoordelen vervult de volgende functies:

- terugkoppeling naar de student;
- terugkoppeling naar de staf over het functioneren van het onderwijs;
- waarmerken van de bekwaamheid der studenten;
- verzekeren van de kwaliteit van de tandheelkundige hulp in de kliniek.

Klinische competentie omvat meer dan het beheersen van motorische vaardigheden. Ook cognitieve en affectieve vaardigheden zijn onontbeerlijk voor goede patiëntenbehandeling. In dit deel van de dissertatie, echter, staat de motorische component centraal.

1.3.2 Specifieke problemen bij het beoordelen van motorische prestaties

Voor het evalueren van motorische vaardigheden zijn "paper-and-pencil tests" niet geschikt. Een student die op papier een uitstekend verhaal kan schrijven over een bepaalde vaardigheid, kan bij de uitvoering ervan op serieuze problemen stuiten. Gelukkig brengen veel motorische vaardigheden produkten voort die geëvalueerd kunnen worden als bewijs van vaardigheid. Evaluatie van motorische vaardigheden vindt dan ook veelal plaats door middel van zogenaamde "work-sample tests". Bij dit soort tests vervaardigen studenten werkstukken waaraan het functioneren in een bepaalde werksituatie (of onderdelen daarvan) getoetst kan worden. Het grote voordeel van work-sample tests is, dat ze direct de betrokken vaardigheden meten en op grond daarvan gekenmerkt worden door een hoge validiteit. Met validiteit wordt dan de inhoudsvaliditeit bedoeld. Bij dit type validiteit is er geen extern criterium waaraan de test gerelateerd kan worden; het criterium is intern. Een test heeft inhoudsvaliditeit als hij een representatieve steekproef, of keuze is van alle mogelijke items die men uit een bepaald gebied kan vormen, terwijl dit gebied

tevens het doel van het testonderzoek is. De representatie is van het externe criterium naar de test zelf verschoven. Het is duidelijk dat inhoudsvaliditeit een kwestie van expert-beoordeling is. Een der gevaren is, dat men daarbij te veel afgaat op de uiterlijke validiteit (face-validity). Bij face-validity geeft de oppervlakkige gelijkenis van de inhoud met wat men wil onderzoeken de doorslag bij de beoordeling, of de test als valide moet worden beschouwd (De Zeeuw, 1978). Het genoegen nemen met face-validity bergt het gevaar in zich dat andere belangrijke test-eigenschappen, zoals betrouwbaarheid, onvoldoende aandacht krijgen.
In verreweg de meeste gevallen is objectief "meten" van de prestatie te verkiezen boven subjectieve evaluatie. Vaak zijn er objectieve aanwijzingen voor de nauwkeurigheid waarmee een student motorische taken heeft uitgevoerd. Op het terrein van de objectieve prestatiemeting wordt onderscheid gemaakt tussen discrete en continue metingen. Heeft men behoefte aan precisie dan moet gekozen worden voor continue metingen. In sommige gevallen is objectieve meting niet mogelijk; bijvoorbeeld als het produkt zich niet leent voor fysieke meting. In dergelijke gevallen kan het beste een aantal variabelen van de resulterende prestatie op een beoordelingsschaal (rating scale) gescoord worden. Aangezien subjectieve evaluatie van prestaties onderhevig is aan een aantal fouten (zie par. 2.2.2), verdient het aanbeveling om een goed geconstrueerde beoordelingsschaal vergezeld te laten gaan van expliciete instructies, die aangeven waar bij het beoordelen op gelet moet worden. Dit zal de betrouwbaarheid over het algemeen ten goede komen. "Over het algemeen" omdat onbetrouwbaarheid door veel meer factoren veroorzaakt kan worden. Bijvoorbeeld doordat de te evalueren prestatie van dien aard is, dat individuele verschillen te klein zijn om betrouwbaarheid te bereiken. Of doordat de beoordelaar geen goede (volledige, ondubbelzinnige) instructies heeft gekregen of niet geoefend werd in het hanteren van het beoordelingsinstrument.
Het belang van het nastreven van betrouwbare instrumenten moge duidelijk zijn. Onbetrouwbare meetinstrumenten zijn oneerlijk voor de student, ongeschikt om te dienen als terugkoppelingsinstrument voor de staf en ongeschikt als instrument voor de kwaliteitsbewaking.

1.3.3 Het vaststellen van tandheelkundige competentie

In par. 1.3.1 werden enkele functies opgesomd van het beoordelen van klinische competentie. De meetmethode die bij dit beoordelen gehanteerd wordt kan absoluut of relatief zijn. Er is sprake van absoluut meten als de bereikte competentie wordt uitgedrukt in termen van nauwkeurig omschreven doelstellingen. Van relatief meten is sprake als de beschrijving van de bereikte competentie een functie is van de prestaties van studiegenoten. Veelal resulteert laatstgenoemde meetmethode in het geven van cijfers op een schaal van 1 tot en met 10, waarbij 6 het laagste voldoende cijfer is. Wie 6 of hoger scoort toont bekwaamheid en mag doorgaan. Wie lager scoort is niet bekwaam en mag het nog eens proberen.

Bij absoluut meten ligt de nadruk niet op selectie maar op diagnostiek. De informatie die de meting oplevert wordt gebruikt om terugkoppeling te verstrekken aan de staf en de student met het doel het onderwijsleerproces te verbeteren en het bereiken van klinische competentie te versnellen.

Klinische competentie kan onderscheiden worden naar kwaliteit en kwantiteit.

Het kwaliteitsaspect

Klinische competentie beschreven in kwalitatieve termen behelst de gevarieerdheid en de mate van excellentie van de verleende diensten. Mackenzie (1973): "If a dentist does only amalgam restorations when other types of restorations are indicated, the quality of care is less than it should be." Traditioneel wordt klinische competentie in de tandheelkunde vastgesteld door de mate van excellentie te bepalen van een vervaardigd produkt. Veel onderzoekers in het tandheelkundig onderwijs pleiten voor aanvulling met andere evaluatievormen. Mackenzie (1973), bijvoorbeeld, benadrukt de noodzaak van ondubbelzinnige informatie over de specifieke leerproblemen van elke student, ten einde goede beslissingen te kunnen nemen. Het uitsluitend beoordelen van eindprodukten, zonder acht te slaan op de factoren die hebben bijgedragen aan verschillen in kwaliteit, leidt tot nodeloze herhaling van leerprocessen. Als een student onvoldoende presteert bij het vervaardigen van een bepaald soort preparatie, dan kan dit het gevolg zijn van:

1. het onvermogen om relevante aanwijzingen te onderscheiden;
2. het niet bezitten van een adekwaat concept met betrekking tot een acceptabel produkt;
3. het niet beheersen van de vereiste motorische vaardigheden.

Instructeurs zouden tests moeten ontwikkelen waarmee vastgesteld kan worden in welke fase van het verwervingsproces studenten zijn vastgelopen.

Het kwantiteitsaspect

Mackenzie (1973) beschrijft kwantiteit in relatie tot klinische competentie als het aantal keer dat een bepaalde procedure uitgevoerd moet worden om van constante kwaliteit verzekerd te zijn. Omdat het leren in de (pre)kliniek kostbaar en tijdrovend is, is het zinvol om te onderzoeken op welke wijze dit leren efficiënter ingericht kan worden. Op welke wijze kan het aantal herhalingen, noodzakelijk voor het leren beheersen van bepaalde procedures, beperkt worden? Mackenzie (1973) noemt een viertal benaderingen voor de oplossing van dit probleem:

1. Het identificeren van gemeenschappelijke componenten in verschillende procedures. Zo is het omgaan met amalgaam hetzelfde voor een aantal amalgaampreparaties. Wordt dit door een student beheerst, dan zou die hiervan vrijgesteld moeten worden. Er resteert dan meer tijd voor het oefenen van andersoortige preparaties.
2. Het gebruik maken van theoretische inzichten in de verwerving van vaardigheden. In par. 1.2 werd uiteengezet hoe motorische

vaardigheden worden verworven. Door met behulp van tests na te gaan of aan de eisen van de leercomponenten wordt voldaan, wordt zinvolle informatie verkregen over de sterke en zwakke punten van individuele studenten. Meer gerichte hulp kan het aantal herhalingen, nodig om competentie te bereiken, reduceren.

3. Het verbeteren van tests, zodat instructeurs sneller competentie kunnen vaststellen. Door gebruik te maken van diagnostische tests kan het aantal onbekende factoren dat variabiliteit veroorzaakt in het eindprodukt, gereduceerd worden. Minder variabiliteit in prestatie resulteert in stabielere prestatiemetingen, waardoor sneller schattingen van het bekwaamheidsniveau van de student gemaakt kunnen worden.
4. Het aantal herhalingen kan ook gereduceerd worden door de beslissingen over individuele studenten te verbeteren. Er zijn twee stappen die het beslissingsproces met betrekking tot het bepalen van de benodigde hoeveelheid oefening, kunnen optimaliseren.
   a. De taken moeten geclassificeerd worden naar moeilijkheidsgraad. Factoren die hierin een belangrijke rol spelen zijn:
      - de ernst van de gevolgen van een gemaakte fout;
      - de mate waarin van routineprocedures wordt afgeweken;
      - de frequentie van voorkomen in de praktijk;
      - het belang van nauwkeurige timing;
      - het belang van snelle uitvoering;
      - de mate waarin precisie vereist is.

      Uiteraard is daarmee de hoeveelheid oefening, nodig voor het bereiken van beheersing, nog niet bepaald. Het leertempo verschilt immers per student. Maar in de toekomst is het wellicht mogelijk om, op basis van leercurves voor specifieke taken, beslissingen te nemen over de hoeveelheid benodigde oefening.
   b. Belangrijker dan de eerste stap is de ontwikkeling van heldere, ondubbelzinnige, betrouwbare en valide prestatiecriteria waarmee de klinische competentie van studenten bepaald kan worden.

In de volgende hoofdstukken van deel I van dit proefschrift wordt beschreven welke activiteiten zijn ondernomen op het terrein van de evaluatie, om het leren in de prekliniek efficiënter te laten verlopen.

II BEOORDELEN VAN MOTORISCHE VAARDIGHEDEN: PROBLEEMSTELLING

2.1 Inleiding

In dit hoofdstuk wordt nader ingegaan op de wijze waarop in het tandheelkundig onderwijs motorische vaardigheden worden vastgesteld. Dergelijke vaardigheden kunnen niet geëvalueerd worden met behulp van paper-and-pencil tests. Meestal wordt geprobeerd om uit de kwaliteit van het vervaardigde produkt af te leiden of de maker ervan voldoende vaardigheid bezit. Als het te beoordelen produkt vervaardigd wordt in een (gesimuleerde) beroepssituatie, dan wordt van "work-sample tests" gesproken. Over de problemen die inherent zijn aan het gebruik van work-sample tests gaat dit hoofdstuk.

2.2 Work-sample tests

2.2.1 Inleiding

Een work-sample test is een replicatie van een werksituatie of onderdeel daarvan en levert metingen op aan de hand waarvan beslissingen genomen kunnen worden over het onderwijs en/of de geëxamineerde. Ze kunnen worden afgenomen in de context van het werk, met gebruikmaking van de reguliere uitrusting en ruimte, of onder simulatie-omstandigheden.
In de geschiedenis van de toegepaste psychologie komt men de work-sample test al vroeg tegen. Wilson (1962) bespreekt een work-sample test voor trambestuurders, ontwikkeld door Munsterberg in het begin van de twintigste eeuw. De belangrijkste toepassing van work-sample tests betrof en betreft de selectie van personeel. Maar, tot op zekere hoogte worden ze ook ingezet voor trainingsdoeleinden.
De toegepaste meet-techniek is doorgaans afhankelijk van de taak die uitgevoerd moet worden. Bestaat de taak uit het omgaan met machines of instrumenten, dan wordt veelal een observator gebruikt die op een check list de waargenomen gedragselementen aankruist. Levert het uitvoeren van de taak een tastbaar produkt op, dan ligt het voor de hand dat de vaardigheid wordt gemeten via dit produkt. In dit verband is men snel geneigd te denken aan industrieprodukten, maar er zijn er veel meer. Ze variëren van eenvoudige handarbeid produkten (bijvoorbeeld een gemetselde muur) tot gecompliceerde esthetische produkten (een schilderij of een muziekstuk). De meetproblemen nemen snel toe naarmate het produkt meer een esthetische component bevat.

2.2.2 Meetproblemen bij het gebruik van work-sample tests

De eis van betrouwbaarheid wordt bij metingen urgent als work-sample tests gebruikt worden om bekwaamheden vast te stellen. De

betrouwbaarheid wordt bedreigd van twee kanten, namelijk van de kant van de geëxamineerde en van de kant van de beoordelaar. Pogingen om menselijke prestaties te meten worden vaak gedwarsboomd door het optreden van variabiliteit. Het is een normaal verschijnsel dat de kwaliteit van de prestatie varieert van meting tot meting. De enige juiste manier om betrouwbare prestatiemetingen te krijgen is dus om de geteste vaker dezelfde test af te nemen. De gemiddelde score zal waarschijnlijk dichter bij de "ware score" liggen dan een score gebaseerd op één meting. Onder de ware score wordt verstaan de gemiddelde score, die de geteste zou halen wanneer de test, of een soortgelijke test, onder alle mogelijke omstandigheden door hem zou worden gemaakt, aangenomen dat geen geheugen- of vermoeidheidsverschijnselen zouden optreden (De Groot, 1975). Helaas ontbreekt het vaak aan tijd om personen meerdere keren dezelfde test af te nemen. In de praktijk wordt meestal met één meting volstaan. In het onderwijs wordt dit onrecht voor de geëxamineerde enigszins verlicht door herkansingen aan te bieden. De gevolgen van onterecht positieve uitslagen, echter, moeten voor lief genomen worden, aangezien het wel haast onmogelijk is om te identificeren wie onterecht geslaagd is.
De tweede bedreiging van de betrouwbaarheid is gelegen in de beoordeling van de prestaties. Of het nu gaat om een proces of een produkt dat beoordeeld moet worden, feit is dat de kwaliteit ervan vaak niet objectief vastgesteld kan worden. Of zoals De Groot (1971) het zegt: "zonder dat de kwalitatieve beschrijving door subjectiviteit kan worden gestoord". Volgens De Groot is een beoordeling pas objectief als er geen "subject" in de zin van een menselijke beoordelaar aan te pas behoeft te komen. Men gaat pas tot het inschakelen van beoordelaars over, omdat men geen betere oplossing weet. Het is dan zaak om de mate van objectiviteit waarmee de beoordelaar te werk gaat zoveel mogelijk te bevorderen. Dit kan geverifieerd worden door de betrouwbaarheid van zijn beoordelingen na te gaan bij een onafhankelijke herhaling van de procedure. Daarnaast, echter, moet ook meer zekerheid verkregen worden over het door de beoordelaar gehanteerde systeem zelf; dat mag ook niet (te) subjectief zijn. Controle daarop kan geschieden door zijn oordeel te vergelijken met dat van andere beoordelaars. Het verdient aanbeveling deze intersubjectieve overeenstemming (inter-judge reliability) empirisch te bepalen. De Groot (1971): "In feite is het vooral dit intersubjectiviteitscriterium, dat bij inschakeling van beoordelaars in de plaats komt van de objectiviteitseis. Qua inhoud zijn de beide begrippen niet gelijkwaardig: volstrekte intersubjectiviteit tussen beoordelaars is (nog) geen objectiviteit, want het systeem is (nog) niet gespecificeerd. Qua strekking zijn de begrippen echter wel zeer verwant. De sociale betekenis van de objectiviteitseis in de wetenschap is immers grotendeels gelegen in het feit, dat waar objectiviteit bestaat volstrekte intersubjectiviteit bereikbaar is; men kan misverstand uitsluiten. Vandaar dat men soms kan volstaan met 'een redelijke mate van intersubjectieve overeenstemming' tussen de tot oordelen bevoegd geachten."
Als vaardigheden niet objectief vastgesteld kunnen worden doet men er dus goed aan om meerdere beoordelaars in te schakelen. Een

voorbeeld moge verduidelijken wat het effect daarvan is op de betrouwbaarheid (gedefinieerd als inter-beoordelaarsovereen - stemming). Sanders (1980) liet 22 tandheelkundige practicum-werkstukken beoordelen door acht, onafhankelijk van elkaar werkende, beoordelaars. Elk werkstuk werd op zes aspecten beoordeeld. In Tabel 2.1 staan de berekende correlatie-coëfficiënten per aspect voor vijf verschillende aantallen beoordelaars. De correlaties zijn Intraklasse correlatie-coëfficiënten (zie par. 5.3.3).

Tabel 2.1: Verband tussen betrouwbaarheid (uitgedrukt in Intraklasse correlatie coëfficiënten) en het aantal beoordelaars.

	aantal beoordelaars				
beoord.asp.	1	2	3	4	8
1	.25	.40	.50	.57	.73
2	.29	.45	.55	.62	.77
3	.22	.36	.46	.53	.69
4	.07	.13	.19	.24	.38
5	.25	.40	.50	.57	.72
6	.30	.46	.56	.63	.77

Uit tabel 2.1 blijkt dat de betrouwbaarheid snel toenam met het inschakelen van grotere aantallen beoordelaars. Echt aanvaardbare waarden werden niettemin pas bereikt toen het aantal beoordelaars gestegen was tot acht. Het inschakelen van meerdere beoordelaars stuit echter dikwijls op praktische bezwaren. Meer dan twee beoordelaars is veelal niet haalbaar in verband met tijdgebrek en/of beschikbaarheid. Ten einde bij een enkel- of tweevoudige beoordeling toch een redelijk vertrouwen te kunnen hebben in de beoordeling, moeten enkele voorzorgen worden genomen die de invloed van de subjectiviteit kunnen beperken. Daarvoor is het eerst nodig om kennis te nemen van de specifieke moeilijkheden, die zich kunnen voordoen bij beoordelingen. De Groot (1971) noemt vijf zogenaamde beoordelaarseffecten:

1. signifisch-effect;
2. halo-effect;
3. sequentie-effect;
4. persoonlijke vergelijkings-effect;

5. contaminatie-effect in engere zin.
Met het signifisch-effect wordt het verschijnsel bedoeld van de beïnvloeding van de beoordeling door de opvatting van de beoordelaar over de beoordelingstaak. Als twee beoordelaars bijvoorbeeld een werkstuk moeten beoordelen op "functionele vormgeving", dan zal de overeenstemming tussen hun oordelen mede afhangen van de vraag of ze beiden hetzelfde verstaan onder "functionele vormgeving".
Het halo-effect treedt op als goede of slechte aspecten van het vervaardigde werkstuk de beoordeling van de andere aspecten in respectievelijk positieve en negatieve zin beïnvloeden.
Als beoordelaars meerdere werkstukken achter elkaar moeten beoordelen, dan kan zich een sequentie-effect voordoen. De beoordeling van een werkstuk wordt dan beïnvloed door voorafgaande werkstukbeoordelingen. Na een aantal slechte werkstukken is men sneller geneigd om een iets beter werkstuk extra positief te beoordelen en omgekeerd.
Met het persoonlijke vergelijkings-effect worden algemeen menselijke, maar ook persoonlijke, neigingen bedoeld tot specifieke beoordelingsverdelingen. Sommige beoordelaars mijden, bijvoorbeeld, extreme beoordelingen (neiging tot het gemiddelde). Daarnaast komt het voor dat beoordelaars consequent te gunstig of te ongunstig beoordelen.
Het contaminatie-effect in engere zin heeft betrekking op het verschijnsel dat een beoordelingsprocedure bewust of onbewust voor andere doeleinden wordt gebruikt dan de beoordeling van de kwaliteit van het werkstuk. Bijvoorbeeld als een beoordelaar zijn studenten eens schrik wil aanjagen omdat hij vindt dat ze hun taak te licht opvatten.
Voor de vijf hierboven genoemde beoordelaarseffecten zijn remedies te bedenken, die weliswaar het beoordelingsprobleem niet kunnen oplossen maar toch in ieder geval kunnen helpen verlichten. In het kader van een meer concrete beoordelings-situatie worden in paragraaf 2.4.3 enkele remedies besproken.

2.3 De work-sample test in het tandheelkundig onderwijs

2.3.1 Inleiding

De belangrijke plaats die motorische vaardigheden in het tandheelkundig onderwijs innemen, kwam reeds ter sprake in hoofdstuk I. In paragraaf 1.3 van dat hoofdstuk werd het belang aangegeven van het vaststellen van het niveau van die vaardigheden en werd tevens gesignaleerd dat strikt objectieve beoordeling niet bereikbaar is. De bij werkstuk-beoordelingen optredende meetproblemen, zoals geschetst in paragraaf 2.2.2, zijn eveneens gesignaleerd in de beoordelingen van tandheelkundige (klinische en preklinische) practicumwerkstukken. Paragraaf 2.3.2 bevat enkele conclusies uit een aantal onderzoeken op dit terrein en gaat nader in op pogingen om de subjectiviteit terug te dringen.

2.3.2 Studies naar de betrouwbaarheid van werkstukbeoordelingen in het preklinisch tandheelkundig onderwijs

Toetsinstrumenten moeten naast de eis van betrouwbaarheid aan nog meer eisen voldoen. Twee andere erg belangrijke eigenschappen zijn objectiviteit en validiteit. De drie genoemde eisen houden zeer nauw verband met elkaar. Niet-objectieve tests zullen door het subjectieve element nooit perfecte overeenstemming tussen beoordelaars kunnen bewerkstelligen, evenmin tussen herhaalde beoordelingen van één beoordelaar. De betrouwbaarheid, gedefinieerd als de inter- en intra-beoordelaarsovereenstemming, zal daardoor gering zijn. Hetzelfde geldt voor de validiteit, want de maximaal haalbare validiteit wordt aangegeven door de betrouwbaarheidsindex (= de wortel uit de betrouwbaarheidscoëfficiënt). Dit betekent dat de kwaliteit van beoordelingssystemen voor een groot deel bekend is, als de betrouwbaarheid ervan bepaald is. In de hieronder te bespreken onderzoeken wordt dan ook bijna uitsluitend gesproken over de overeenstemming die beoordelaars onderling of met zichzelf bereiken. Overigens wordt niet gestreefd naar volledigheid bij het bespreken van de onderzoeksliteratuur op het gebied van de tandheelkundige evaluatie. In de eerste plaats omdat voornamelijk onderzoeken worden besproken die zijn uitgevoerd in het preklinisch onderwijs (in verband met het eigen onderzoek, dat zich beperkt tot het beoordelen in het preklinisch practicum). In de tweede plaats omdat lang niet alle studies nieuwe informatie toevoegen aan wat reeds bekend is.

In 1967 rapporteren Natkin en Guild (1967) over de invoering van een zogenaamde "grade card". Cijfers alleen werden onvoldoende betrouwbaar geacht om prestaties te beschrijven. Vandaar dat elk cijfer voortaan vergezeld werd van commentaar, om het gegeven cijfer te motiveren. Vergelijking van grade cards bracht aan het licht dat cijfers vaak inconsistent en willekeurig gegeven werden en (gezien het commentaar) irrationeel waren. Bovendien waren de commentaren niet informatief met betrekking tot de geleverde prestaties. Het identificeren van hiaten in de betreffende vaardigheid bleef daardoor achterwege. De waargenomen inconsistenties bij klinische evaluaties moedigden een meer systematische studie aan in het preklinisch onderwijs, waar de te beoordelen prestaties minder complex zijn en een meer nauwkeurige experimentele controle beter mogelijk is. Zes beoordelaars evalueerden onafhankelijk van elkaar 65, door studenten uitgevoerde, endodontische behandelingen. De resultaten waren onthutsend; voor 45 procent van de werkstukken varieerde het toegekende cijfer over vier of meer cijfers. Slechts voor vijf werkstukken kenden alle beoordelaars cijfers toe die niet meer dan één cijfer van elkaar verschilden. Echter, ondanks identieke of bijna identieke cijferwaarderingen, gaven beoordelaars vaak tegengesteld commentaar als motivering voor het toegekende cijfer. De auteurs trokken de volgende conclusies:

1. Niet alle beoordelaars ontdekken dezelfde fouten. Het komt herhaaldelijk voor, dat een door een beoordelaar gesignaleerde fout niet ontdekt wordt door een andere beoordelaar of niet als

fout wordt aangemerkt.
2. Zelfs wanneer beoordelaars dezelfde tekorten signaleren, kennen ze vaak zeer uiteenlopende "straffen" toe.

Fuller (1972) probeerde vast te stellen of kritiek op de gebruikte beoordelingsmethode gerechtvaardigd was. Hij liet acht medewerkers een aantal van 67, door eerstejaars studenten, vervaardigde werkstukken beoordelen op de traditionele "glance-and-grade" manier. Daarna moesten ze een gestratificeerde random steekproef van 25 werkstukken opnieuw beoordelen om de intra-beoordelaarsbetrouwbaarheid te kunnen bepalen. De correlatie-coëfficiënten, berekend tussen de eerste en de tweede beoordeling, werden gebruikt om beoordelaarsparen samen te stellen; de hoogste en laagste coëfficiënt vormden een paar, de op één na hoogste en de op één na laagste een ander paar, enz. Vervolgens werd voor elk beoordelaarspaar de inter-beoordelaarsbetrouwbaarheid vastgesteld. De Spearman rangcorrelatie coëfficiënten, gebruikt voor de beschrijving van de intra-beoordelaarsbetrouwbaarheid, varieerden van .47 tot .83. De inter-beoordelaarsbetrouwbaarheid werd beschreven door middel van correlatie-coëfficiënten, gebaseerd op variantie-analyses. De coëfficiënten varieerden van .24 tot .53.
Fuller concludeerde dat:
1. er geen sprake was van significante overeenstemming tussen beoordelaars;
2. de stabiliteit van beoordelingen varieerde van acceptabel tot onvoldoende.

Houpt en Kress (1973) lieten instructeurs, studenten en tandarts-assistenten werkstukken beoordelen op globale en analytische (acht criteria) wijze. Beoordelaars waren op willekeurige wijze toegewezen aan drie deelgroepen. Drie verschillende beoordelingsschalen* werden door die deelgroepen gebruikt bij het beoordelen. De intra- beoordelaarsbetrouwbaarheid voor globale beoordelingen was redelijk tot hoog; vijf van de negen betrouwbaarheidsschattingen waren gelijk aan of groter dan .75. Voor de analytische beoordelingen waren 21 van de in totaal 72 betrouwbaarheidsschattingen redelijk tot hoog. De inter-beoordelaarsbetrouwbaarheid voor globale beoordelingen varieerde van .45 tot .88 en was in vijf van de negen gevallen gelijk aan of groter dan .75 (correlatie-coëfficiënten gebaseerd op variantie-analyse). Van de analytische beoordelingen bleken slechts 5 van de 72 schattingen een redelijke tot hoge waarde te hebben. Over het algemeen waren de verschillen in betrouwbaarheid tussen instructeurs en studenten niet significant maar tussen instructeurs en tandarts-assistenten en studenten en tandarts-assistenten wel.

*Een tweepunts-schaal zonder gespecificeerde schaalpunten;
een vijfpunts-schaal met gespecificeerde eindpunten;
een vijfpunts-schaal met gespecificeerde schaalpunten.

Met betrekking tot de betrouwbaarheid waren de verschillen tussen de groepen, die de verschillende schalen gebruikten, gering. De nauwkeurigheid (gedefinieerd als de overeenstemming met expert-oordelen) was echter duidelijk gebaat met het gebruik van een tweepunts-schaal.
Conclusies:
1. Als werkstukken niet op totaalniveau beoordeeld worden maar op een aantal relevante aspecten (criteria), dan zijn de beoordelingen niet betrouwbaar.
2. Vakkennis is geen strikte noodzaak om beoordelaars nauwkeurig te laten beoordelen.
3. Tweepunts-schalen leveren meer nauwkeurige beoordelingen op dan vijfpunts-schalen.

Vele andere studies werden verricht met (ten dele) andere vraagstellingen en andere onderzoeksmethoden. Zonder uitzondering kwamen ze voort uit onvrede met de vigerende beoordelingssystemen. Sommige studies, zoals de hierboven behandelde, illustreerden de problemen aan de hand van empirisch materiaal. Allemaal, echter, waren ze erop gericht om de beoordelingskwaliteit te verbeteren. Die pogingen spitsten zich toe op de volgende drie aspecten:
a. prestatiecriteria;
b. scoring;
c. training van beoordelaars.

ad a. Prestatiecriteria

Men kan proberen de subjectiviteit in beoordelingen terug te dringen door zeer nauwkeurig de eisen te formuleren waaraan het werkstuk moet voldoen. Door de complexiteit van tandheelkundige werkstukken kan dit niet voor totaal-beoordelingen gerealiseerd worden, zodat voor betrouwbare beoordelingen meerdere prestatiecriteria afgeleid moeten worden. Belangrijke vragen in dit verband zijn:
- Hoe moeten prestatiecriteria afgeleid worden?
- Aan welke eisen moeten ze voldoen?

Patridge en Mast (1978) schrijven in een overzichtsartikel over tandheelkundige evaluatie, dat hen geen onderzoek bekend is, waarin verschillende procedures voor het afleiden van prestatiecriteria met elkaar vergeleken worden. De auteurs maken melding van een drietal verschillende methodes:
- Een methode die voornamelijk berust op het bestuderen van wat anderen gedaan hebben op hetzelfde gebied. Aanwijzingen hierover kunnen gevonden worden in instructieboeken, in gehanteerde beoordelingsmethoden en in de expertbeoordelingen van de doelvaardigheid.
- Een statistische methode. Hierbij wordt gekeken naar de hoeveelheid variantie die prestatiecriteria kunnen verklaren in de totaalscore. Alleen de prestatiecriteria die een flink deel van de variantie verklaren, moeten in het beoordelingssysteem worden opgenomen.
- Het zelfstandig uitvoeren van een taakanalyse. Een taakanalyse

is het ontleden van een taak in zinvolle delen en het beschrijven van de relatie tussen die delen. Op deze wijze worden de cruciale aspecten van een vaardigheid geïdentificeerd.

De eisen waaraan prestatiecriteria moeten voldoen liggen voor de hand; dié maatregelen dienen te worden genomen die de betrouwbaarheid en validiteit bevorderen. Mackenzie (1974) noemt een vijftal eisen:

1. Criteria moeten direct gerelateerd zijn aan de doelstellingen. Dat wil zeggen dat de nadruk moet liggen op dié aspecten die direct van invloed zijn op de kwaliteit van het werkstuk.
2. Criteria dienen operationeel gedefinieerd te worden. Dat wil zeggen dat ze in meetbare termen omschreven moeten worden.
3. Criteria moeten vermelden wat nog wel en wat niet meer acceptabel is en, bij gebruik van meerpunts-schalen, de grenzen van de klassen, zodat beoordelaars een werkstuk eenvoudiger kunnen classificeren.
4. Criteria moeten zo weinig mogelijk categorieën definiëren. Alleen dié categorieën moeten gedefinieerd worden, die leiden tot nuttige beslissingen en acties.
5. De omschrijvingen moeten verduidelijkt worden met behulp van een illustratie of een model.

In verschillende experimentele studies (Gaines, Rasmussen en Uchello, 1975; Natkin en Guild, 1967; Abou-Rass, 1973) werd gewerkt met nauwkeurig gedefinieerde prestatiecriteria. Jammer genoeg werden tegelijkertijd ook andere onafhankelijke variabelen ingevoerd, zodat definitieve conclusies over de effecten van prestatiecriteria niet getrokken kunnen worden.

ad b. Scoring

Er zijn twee manieren om prestaties te scoren: globaal en analytisch. Globale scoring benadert de prestatie als geheel; onderdelen van die prestatie worden niet gescoord. Een bekend voorbeeld is de zogenaamde "glance-and-grade" methode. Analytische scoring, daarentegen, levert aparte beoordelingen op voor elk van de onderdelen waaruit de prestatie bestaat. Beide systemen kunnen in principe betrouwbare metingen opleveren. De moeilijkheid bij globale scoring betreft de praktische onmogelijkheid om de eisen waaraan het werkstuk moet voldoen zodanig te omschrijven, dat er geen gevaar bestaat, dat verschillende beoordelaars zich in hun beoordeling laten leiden door verschillende aspecten. De analytische scoring wordt steeds vaker gebruikt. Het grote voordeel hiervan is de verbeterde terugkoppeling naar de student en het relatieve gemak waarmee de prestatiecriteria op objectieve wijze geformuleerd kunnen worden. Dit laatste omdat men zich bij de omschrijving volledig kan richten op dat ene deelgebied dat beoordeeld moet worden.
Een belangrijke beslissing bij de constructie van een beoordelingsinstrument betreft de keuze van het aantal schaalpunten van de beoordelingsschaal. De eenvoudigste beoordelingsschaal kent twee schaalpunten en kan opgevat worden als een goed/fout-schaal;

het werkstuk voldoet wel of niet aan de eisen. Er van uitgaande dat de grens tussen goed en fout op vrij ondubbelzinnige wijze te omschrijven is, moeten dergelijke schalen een vrij hoge overeenstemming tussen beoordelaars kunnen bewerkstelligen. Daar tegenover staan enkele nadelen:

- Tandheelkundige vaardigheden zijn vaker niet dan wel goed/fout-vaardigheden. In veel gevallen is het mogelijk om werkstukken te maken waarvan de te onderscheiden aspecten niet als ideaal beoordeeld kunnen worden, maar wel klinisch acceptabel zijn.
- Voor terugkoppelingsdoeleinden is een tweepunts-schaal niet erg geschikt. De efficiëntie van het onderwijsleerproces is gebaat bij het verstrekken van zodanige terugkoppeling aan de lerende, dat deze op grond van de beoordeling weet wat hem te doen staat. Nodeloze herhaling wordt daarmee voorkomen.

In instructieve situaties zijn meerpunts-schalen dus geschikter dan tweepunts-schalen. Helaas wordt het objectief formuleren van prestatiecriteria steeds moeilijker als het aantal schaalpunten toeneemt. Het verschil tussen "goed" en "zeer goed" bijvoorbeeld, laat zich minder makkelijk beschrijven dan het verschil tussen "voldoende" en "onvoldoende". Bovendien zijn steeds langere omschrijvingen nodig naarmate fijnere differentiaties gemaakt moeten worden, waardoor de hanteerbaarheid van het beoordelingsinstrument bedreigd wordt.

In diverse studies werd de relatie tussen het aantal schaalpunten en de beoordelaarsovereenstemming onderzocht. Houpt en Kress (1973) bijvoorbeeld, vonden dat zowel intra- als inter-beoordelaarsovereenstemmingen hoger waren bij gebruik van een tweepunts-schaal dan bij een vijfpunts-schaal. Hinkelman en Long (1973) vergeleken een driepunts-schaal met een tweepunts-schaal en vonden dat de overeenstemming tussen beoordelaars ongeveer tien procent groter was bij gebruik van de tweepunts-schaal. Vermeld dient te worden dat er geen echte scoring op tweepunts-schalen had plaatsgevonden. Door de categorieën "geen verbeteringen nodig" en "klinisch acceptabel" onder te brengen in een "geslaagd" categorie, werden de scores op de driepunts-schaal getransformeerd naar een tweepunts-schaal. Het samenvoegen van categorieën kan echter alleen maar leiden tot hogere of gelijke (in het geval van niet gebruikte categorieën) overeenstemming tussen beoordelaars. Conclusies ten gunste van tweepunts-schalen zijn op grond van deze studie derhalve onmogelijk. Dat meerpunts-schalen ook betrouwbaar kunnen zijn demonstreerden Ryge en Snyder (1973). Hun vierpunts-schaal werd door speciaal geselecteerde en getrainde beoordelaars gebruikt bij het beoordelen van een kleine duizend restauraties. De overeenstemming tussen twee beoordelaars bedroeg 92 procent en de intra-beoordelaarsovereenstemming 89 procent. Helaas hadden de auteurs de restauraties niet ook nog op een tweepunts-schaal laten beoordelen door dezelfde beoordelaars, waardoor een uitspraak over de relatie tussen het aantal schaalpunten en de betrouwbaarheid onmogelijk werd.

Conclusies over het optimale aantal schaalpunten voor het beoordelen van tandheelkundige practicumwerkstukken kunnen, op basis van bovenstaande onderzoeken, niet getrokken worden. Getwijfeld moet worden aan het bestaan van zo'n optimum.

Het lijkt verstandiger om het aantal schaalpunten afhankelijk te laten zijn van het doel van de beoordeling en van de vaardigheid die beoordeeld wordt. Is het doel van de beoordeling om na te gaan wie de vaardigheid beheerst en wie nog niet, dan is een tweepunts-schaal te verkiezen boven een meerpunts-schaal. Het doel van het beoordelen is dan selectie. Is het doel meer gericht op diagnose, dat wil zeggen op het ontdekken van tekortkomingen in de vaardigheid, dan kan beter met een meerpunts-schaal gewerkt worden. De student kan in dat geval een meer gerichte terugkoppeling verwachten. Eveneens gaat de voorkeur uit naar meerpunts-schalen als de vast te stellen vaardigheid zich leent voor het maken van zinvolle onderscheidingen in prestaties. Zinvolle verschillen zijn verschillen die tot uiting komen in de duurzaamheid van het behandelingsresultaat (bijvoorbeeld de levensduur van een gelegde vulling), het comfort van de patiënt en in het esthetische aspect.

Een probleem van deelscores bij analytische scoring is het nemen van zak-/slaagbeslissingen. Op welke wijze moet de informatie die de deelscores opleveren gebruikt worden om te bepalen wie de vaardigheid voldoende beheerst en wie niet? Als alle gescoorde onderdelen even belangrijk zijn, kunnen zak-/slaagbeslissingen eenvoudig genomen worden op basis van een percentage; bijvoorbeeld 80 procent van de scores moet voldoende zijn. Zijn niet alle onderdelen even belangrijk, wat meestal het geval zal zijn, dan kunnen de deelscores gewogen worden, zodat het relatieve belang tot uiting komt in de totaalscore. Overigens wordt aan het nut van weging voor de betrouwbaarheid en validiteit getwijfeld. Correlaties tussen wel en niet gewogen totaalscores op cognitieve tests zijn zeer hoog (Fitzpatrick en Morrison, 1971). Het toekennen van gewichten aan diverse deelaspecten gebeurt meestal op vrij arbitraire wijze: een docent of een groep docenten vindt, op grond van ervaring en/of studie, dat het ene aspect meer bijdraagt aan het totaalresultaat dan een ander aspect. Verreweg de beste methode om de gewichten van de diverse aspecten vast te stellen, is door middel van systematisch onderzoek na te gaan welke onderdelen van een vaardigheid het belangrijkste zijn voor de beheersing ervan. Weinig is bekend over de effecten van weging op de beoordeling van tandheelkundige vaardigheden. Gaines, Rasmussen en Uchello (1975) kenden verschillende gewichten toe aan prestatiecriteria. In een door hen uitgevoerd vergelijkend onderzoek tussen een glance-and-grade methode en de methode met de gewogen prestatiecriteria was de inter-beoordelaarsbetrouwbaarheid respectievelijk 0.03 en 0.65. Door de opzet van de studie was het niet mogelijk om vast te stellen of de verbetering veroorzaakt was door het gebruik van prestatiecriteria, gewogen scoring, training, een driepunts-schaal, of een combinatie van deze variabelen.

ad c. Training van beoordelaars

De kans dat beoordelingsschalen betrouwbare informatie opleveren wordt groter als de opvatting van de beoordelingstaak gelijk is voor alle beoordelaars. Vaak blijkt dat mondelinge en/of schriftelijke instructies verschillen in taakopvatting niet kunnen

voorkomen. Selectie van beoordelaars lijkt een voor de hand liggende oplossing. Onderzoek zou bijvoorbeeld kunnen uitwijzen welke beoordelaars de taak op dezelfde wijze interpreteren. Bezwaren tegen deze oplossing zijn:

1. Meestal zijn er niet zo veel beoordelaars dat men zich de luxe van selectie kan permitteren.
2. Veelal maakt de beoordelingstaak deel uit van het werk; ontkoppeling kan dan worden opgevat als een motie van wantrouwen.

Een andere mogelijkheid om interpretatieverschillen van de beoordelingstaak terug te dringen, is training van beoordelaars. Johnson (1972) noemt verschillende onderzoeken op psychologisch terrein, waarin training van beoordelaars de nauwkeurigheid van beoordelingen positief beïnvloedde. De invloed van training werd ook onderzocht in een aantal studies naar de kwaliteit van werkstukbeoordelingen in het tandheelkundig onderwijs.

Natkin en Guild (1967) bijvoorbeeld, lieten vijf beoordelaars deelnemen aan een drietal studies. In studie 1 moesten deze beoordelaars aan 50 elementen een cijfer toekennen en dit becommentariëren. In studie 2 werden 25 elementen beoordeeld aan de hand van een nieuw beoordelingssysteem, waarmee de beoordelaars door lezen hadden kennis gemaakt. Studie 3 was identiek aan studie 2, maar vond pas plaats nadat acht trainings-sessies waren gehouden. In elke trainings-sessie werden tien werkstukken beoordeeld aan de hand van de nieuwe beoordelingsmethode. Na afloop van elke sessie werden de resultaten besproken, met het doel beoordelaars vertrouwd te maken met methodes en criteria voor het herkennen en categoriseren van fouten. De grootste toename in overeenstemming trad op bij beoordelaars die door lezen hadden kennis genomen van de beoordelingsmethode. Training bleek geen significant effect te hebben op de inter-beoordelaarsovereenstemming.

Fuller (1972) liet een aantal beoordelaars deelnemen aan een trainingsprogramma dat éénmalig was en bestond uit de volgende onderdelen:

1. Begrippen en technieken die aan studenten werden geleerd, werden verklaard en besproken met gebruikmaking van modellen.
2. Menselijke factoren die het beoordelen kunnen beïnvloeden werden besproken.
3. De gebruikte beoordelingsmethode werd uitvoerig uitgelegd.
4. De beoordelingstaak werd geoefend. De onderzoeker vergeleek zijn eigen beoordelingen met die van de trainees en besprak eventuele discrepanties.

Fuller vond geen enkel bewijs dat de betrouwbaarheid van de beoordelingen toenam als gevolg van training.

Abou-Rass (1973) onderscheidde in de endodontische behandeling een aantal stappen en construeerde op basis daarvan twee beoordelingsinstrumenten. Het tweede instrument was een verfijning van het eerste. Beoordelaars werden getraind in het gebruik van het tweede instrument. De training bestond uit zes zittingen, waarin voor elke "stap" de betekenis werd besproken voor de

endodontische procedure, de meetmethode en de evaluatie van het gemetene. De overeenstemming tussen de beoordelaars steeg van 54 procent vóór de training tot 77 procent erna.

Hinkelman en Long (1973) lieten vier stafleden onafhankelijk van elkaar 30 preparaties beoordelen, gebruik makend van een nieuw beoordelingssysteem. Na een week beoordeelden dezelfde vier personen 30 andere preparaties. In de periode tussen de twee beoordelingsronden werd een trainings-sessie gehouden om de beoordelaars meer vertrouwd te maken met de nieuwe beoordelingsmethode. Slechts een geringe verbetering in de overeenstemming werd gevonden. Niettemin zagen de onderzoekers daarin een reden om met enige regelmaat trainings-sessies te organiseren. Op die wijze hoopten zij uniforme interpretatie te bereiken én te handhaven.

Houpt en Kress (1973) beschrijven de resultaten van een onderzoek waarin training omschreven wordt als het geven van onmiddellijke terugkoppeling aan beoordelaars over hun beoordelingsprestaties. De training had een gering positief effect op de intra-beoordelaarsovereenstemming bij tandarts-assistenten, maar niet bij studenten en stafleden. Ook werden geen significante effecten gevonden met betrekking tot de inter-beoordelaarsovereenstemming.

In Nederland ontwikkelden Steures en Tromp (1980) "calibratietrainingen" ten behoeve van de docenten van het preklinisch practicum. Als voornaamste doelen werden genoemd: het vaststellen van verschillen tussen docenten in de beoordeling van de prestaties van de studenten; bespreking van de mogelijke oorzaken en terugdringen van deze verschillen. De auteurs schetsen het verloop van een calibratietraining onder vermelding dat die "model" staat voor het merendeel van de in de loop der tijd gehouden trainingen. De training bestond uit twee bijeenkomsten met een week tussentijd. Dertien beoordelaars (11 student-instructeurs en 2 tandartsen) beoordeelden een aantal werkstukken op veertien criteria. De verschillen in beoordeling werden spectaculair genoemd. Het doel van de tweede bijeenkomst was om na te gaan in hoeverre de geconstateerde verschillen in omvang waren terug gebracht. Acht docenten waren milder geworden in hun oordeel, vijf strenger. De tweede keer werden meer criteria milder beoordeeld dan de eerste keer. Helaas besteedden de auteurs in hun artikel geen aandacht aan de intra- en inter- beoordelaarsovereenstemming, zodat geen effect vastgesteld kon worden van training op de betrouwbaarheid van beoordelingen.

Uit de aangehaalde literatuur blijkt hoe inconsistent de resultaten zijn van het onderzoek naar de effecten van training op de betrouwbaarheid van werkstukbeoordelingen. Slechts één studie rapporteert duidelijke winst. Deze wat teleurstellende resultaten mogen echter niet leiden tot de uitspraak, dat training van beoordelaars niet kan helpen om werkstukbeoordelingen betrouwbaarder te laten zijn. Daarvoor verschillen de besproken onderzoeken te veel van elkaar in opzet en is te weinig bekend over de

wijze waarop trainings-sessies werden ingericht.

2.4 De beoordelingskwaliteit van preklinische werkstukken in de Nijmeegse Subfaculteit Tandheelkunde

2.4.1 Inleiding

In 1981 werd in Nijmegen een rapport uitgebracht van de Subcommissie Toetsing en Beoordeling Motorische Vaardigheden (Otto, 1981). Het rapport beoogde een overzicht te geven van de gehanteerde beoordelingsprocedures in de motorische onderwijsblokken van het eerste cursusjaar Tandheelkunde. Tevens werd een inventarisatie gemaakt van verricht onderzoek naar de beoordelingskwaliteit in die onderwijsblokken. Op grond van die gegevens deed de Subcommissie enkele aanbevelingen voor de verbetering van beoordelingsprocedures.
Achtereenvolgens komen in de volgende subparagrafen aan de orde:
- De resultaten van enkele onderzoeken met betrekking tot de beoordelingskwaliteit in de motorische onderwijsblokken (par. 2.4.2).
- Voorstellen voor het verbeteren van beoordelingsprocedures (par. 2.4.3).

2.4.2 Resultaten van enkele subfacultaire onderzoeken naar de beoordelingskwaliteit van motorische vaardigheden

In 1973 analyseerde Borgesius (1973) de beoordelingen van preklinische practicumwerkstukken (preparaties en restauraties) van eerste- en tweedejaars studenten. Vier verschillende typen werkstukken werden op een goed/fout schaal beoordeeld door vijf, onafhankelijk van elkaar werkende, beoordelaars. Hij berekende de inter-beoordelaarsovereenstemmingen voor alle mogelijke beoordelaarsparen op totaalniveau (som van de itemscores) en op itemniveau. Als overeenstemmingsmaat werd Pearson's correlatie coëfficiënt gebruikt. De gemiddelde correlatie coëfficiënt op totaalniveau bedroeg 0.36, op itemniveau 0.15.

Otto (1979a) bestudeerde de beoordeling van opwaswerkstukken in het eerstejaarspracticum "contourherstel van solitaire elementen" (blok 160c). De beoordeling in dit blok geschiedde aan de hand van zes goed/fout items door drie beoordelaars. Via een stemmingsregeling werden de drie beoordelingen per item omgezet in één oordeel per item en vervolgens getransformeerd naar een cijfer op een schaal van 1 tot 9. Voor zijn onderzoek transformeerde Otto de itemscores van elke beoordelaar naar een cijfer tussen 0 en 8. Voor drie beoordelaarsparen werden Pearson correlatie coëfficiënten berekend over de resulterende cijfers. De correlaties waren redelijk (0.74), matig (0.49) en laag (0.31). Ook de zak-/slaagbeslissingen werden bestudeerd. Bij één beoordelaarspaar werden geen verschillen in afwijzingsproporties aangetroffen. Bij een ander paar was er sprake van een groot verschil: 0.95 versus 0.41.

Het laatste paar kende zeer weinig verschil: 0.61 versus 0.64. Vergelijking tussen de beoordelaarsparen was niet mogelijk omdat de drie paren verschillende werkstukken beoordeelden.

Sanders (1980) liet in het eerstejaars practicum "prepareren en restaureren" (blok 155) acht beoordelaars een serie van 22 preparatiewerkstukken (klasse II) beoordelen op zes items. De items werden gescoord op een driepunts-schaal: 1 = de kwaliteit is slecht; 2 = de kwaliteit voldoet niet geheel aan de eisen, maar is acceptabel; 3 = de kwaliteit is uitstekend.
In zijn onderzoek telde Sanders de itemscores op om totaalscores te verkrijgen. Over de totaalscores werden inter-beoordelaarsovereenstemmingen berekend voor alle mogelijke beoordelaarsparen. De gebruikte overeenstemmingsmaat was de Pearson correlatie coëfficiënt. De 28 resulterende coëfficiënten varieerden in grootte van 0.18 (lage overeenstemming) tot 0.65 (matige overeenstemming). Met gebruik making van de in het onderwijs gehanteerde transformatieregel werden de scores omgezet in cijfers op een tienpunts-schaal. Vervolgens kon de consistentie van zak-/slaagbeslissingen worden nagegaan door voor elk beoordelaarspaar een kruistabel op te stellen van voldoende/onvoldoende-beslissingen. De overeenstemmingen werden uitgedrukt in a-waarden (proportie overeenstemming) en ac-waarden (voor kans gecorrigeerde overeenstemmingen). De proportie overeenstemming varieerde van 0.41 (matig) tot 0.95 (hoog) maar lag wel steeds (één geval uitgezonderd) boven de kansovereenstemming, zodat van reële overeenstemming tussen de beoordelaars gesproken kon worden.

Bovengenoemde onderzoeken en nog enkele andere onderzoeken in het eerstejaars practicum "mondhygiëne" (blok 152) (Otto, 1979b) en het eerstejaars practicum "contourherstel gemutileerde dentitie" (blok 164) (Borgesius, 1980) brachten de Subcommissie tot de conclusie dat...."zowel bij het toekennen van cijfers voor, als bij het nemen van voldoende/onvoldoende-beslissingen over werkstukken (c.q. handelingen) de tussenbeoordelaarsovereenstemming op totaalniveau matig tot laag is."

2.4.3 Advies van de Subcommissie voor het verbeteren van beoordelingsprocedures

De Subcommissie onderscheidt een aantal fasen in het construeren van nieuwe beoordelingsprocedures en geeft voor elke fase adviezen, die volgens haar kunnen leiden tot betere beoordelingen. Deze adviezen worden hier onderschreven en opgenomen met het oog op de in hoofdstuk III beschreven ontwikkeling van een beoordelingsmethode voor preklinische practicumwerkstukken.

1. Specificeren van de einddoelstellingen van het blok
   Wil men op grond van beoordelingen uitspraken doen over het al dan niet bereiken van doelstellingen, dan moet er voor gezorgd worden dat de toets zo'n uitspraak mogelijk maakt. Met andere woorden: de toets moet representatief zijn voor de doelstel-

lingen. Vandaar dat het absoluut noodzakelijk is dat in de einddoelstellingen van de onderwijsblokken gespecificeerd wordt welke typen werkstukken gemaakt moeten worden en welke handelingen verricht moeten worden.

2. Constructie van een analytische beoordelingsprocedure
De Subcommissie is van oordeel dat een analytische beoordelingsprocedure de meeste aanknopingspunten biedt voor verbeteringen. Belangrijke punten zijn:
- het opstellen van een itemlijst. Welke aspecten kunnen worden onderscheiden en welke daarvan dienen beoordeeld te worden?
- het formuleren van itemscore-regels. Voor elk item moet omschreven worden aan welke eisen het werkstuk met betrekking tot dat item moet voldoen. Zo mogelijk dient een meetprocedure gespecificeerd te worden waarmee men kan uitmaken welke score in een concreet geval van toepassing is.
- er moeten regels geformuleerd worden voor het combineren van itemscores tot een totaalscore, voor het bepalen van de cesuur en voor het omzetten van totaalscores in cijfers.

3. De feitelijke beoordelingen
De Subcommissie beveelt aan om met meerdere beoordelaars te werken en deze beoordelaars te trainen. Verder moet bijzondere aandacht besteed worden aan de zogeheten beoordelaarseffecten (zie paragraaf 2.2.2). Voor de beschreven effecten worden enkele remedies gegeven.
Om het signifisch effect te verminderen dient de beoordelingstaak voor de beoordelaar zo duidelijk mogelijk omschreven te zijn en mag de beoordelingstaak niet te ingewikkeld zijn. Halo-effecten kunnen gereduceerd worden door er op toe te zien dat beoordelaars niet meer informatie krijgen dan voor het uitvoeren van de beoordelingstaak nodig is. Dit is eenvoudig te realiseren door beoordelaars niet te laten weten van welke student(en) ze het (de) werkstuk(ken) beoordelen. Een andere maatregel voor het reduceren van het halo-effect is minder eenvoudig te realiseren. Bedoeld wordt de eis dat beoordelaars zich volledig op het te beoordelen item concentreren. Opsplitsing in letterlijke zin is bij tandheelkundige werkstukken nu eenmaal onmogelijk, zodat beoordelaars altijd het totale werkstuk voor zich hebben, ook al beoordelen ze slechts één aspect. Een mogelijke oplossing is het bekijken per item of per groep van items van een reeks te beoordelen werkstukken. Voor de student heeft dit tot gevolg dat zijn prestatie niet meer onmiddellijk beoordeeld wordt, wat nadelig kan zijn voor de terugkoppeling.
Sequentie-effecten hebben minder invloed als werkstukken meer dan één keer door dezelfde beoordelaar beoordeeld worden. De volgorde moet dan gewijzigd worden. Een nadeel is dat er veel tijd mee gemoeid is. Van de andere kant levert dit het voordeel op dat intra-beoordelaarsovereenstemmingen berekend kunnen worden.
Voor het verminderen van persoonlijke vergelijkings-effecten heeft de Subcommissie eigenlijk geen echte remedie beschik-

baar. Ze beveelt aan om beoordelingsresultaten te administreren, opdat opvallende beoordelaarskarakteristieken in ieder geval gesignaleerd kunnen worden.
Om het contaminatie-effect in engere zin te verminderen is het nodig om werkstukken te laten beoordelen door meerdere, niet belanghebbende en onafhankelijk werkende beoordelaars. Dat beoordelaars een zeker belang hebben bij hun beoordelingen lijkt nauwelijks te vermijden. Om onafhankelijk beoordelen te garanderen zijn verschillende manieren denkbaar. Erg voor de hand liggend is dat de beoordelaars niet met elkaar in contact kunnen komen als ze met de beoordelingstaak bezig zijn.

4. Het verwerken en analyseren van beoordelingsgegevens
Na beoordeling van een aantal werkstukken beschikt men veelal over een groot aantal beoordelingsgegevens, die volgens vaste procedures verwerkt moeten worden tot cijfers. Het automatiseren hiervan is erg aantrekkelijk in verband met de tijdwinst en de grote nauwkeurigheid waarmee de berekeningen uitgevoerd kunnen worden. Daarnaast biedt automatisering van de verwerking van de beoordelingsgegevens goede aanknopingspunten om het analyseren van deze gegevens ook geautomatiseerd te laten verlopen. De analyse van beoordelingsgegevens is nodig om de kwaliteit van de beoordelingen te kunnen vaststellen. Op grond van de resultaten van een analyse kunnen de beoordelingsprocedure en de uitvoering ervan, indien nodig, worden bijgesteld. Blijkt de betrouwbaarheid van de beoordelingen niet aan van te voren gespecificeerde minima te voldoen, dan kan bijvoorbeeld besloten worden opnieuw te beoordelen, alvorens de cijfers definitief vast te stellen.

2.5 Een onderwijsstimuleringsproject

In september 1981 startte binnen het Instituut Conserverende Tandheelkunde voor Volwassenen een onderwijsstimuleringsproject (osp), dat tot doel had, middels een geïndividualiseerd trainingsprogramma, de beoordelingskwaliteit te verbeteren van preklinische practicumwerkstukken. Hoewel niet noodzakelijk voor het functioneren van het trainingsprogramma werd tegelijkertijd een nieuwe beoordelingsmethode ontwikkeld. Het trainingsprogramma en de nieuwe beoordelingsmethode werden geïntegreerd getest, waarbij tevens geprobeerd werd om een vergelijking te maken tussen de oude en de nieuwe beoordelingsmethode met betrekking tot de beoordelingskwaliteit.
Bij de opzet van het osp is zo veel mogelijk rekening gehouden met de aanbevelingen van de Subcommissie Toetsing en Beoordeling Motorische Vaardigheden, zoals die in par. 2.4.3 zijn besproken. Aan de volgende aanbevelingen van de Subcommissie is tegemoet gekomen:
- er is een itemlijst opgesteld van te onderscheiden beoordelingsaspecten;
- voor elk item wordt het criterium in zo objectief mogelijke termen omschreven;

- voor elk item wordt aangegeven op welke wijze vastgesteld kan worden of aan het criterium voldaan wordt (beoordelingsprocedure);
- beoordelaars worden getraind;
- er is een transformatieregel ontwikkeld om totaalscores om te zetten in cijfers op een schaal van 1 tot 10;
- er is een administratiesysteem opgezet voor het bijhouden van beoordelingsgegevens;
- scoring en administratie zijn geautomatiseerd;
- in het trainingsprogramma wordt de terugkoppeling over de beoordelingsprestaties verzorgd door een microcomputer.

De centrale vraagstelling luidt als volgt: Resulteren bovengenoemde maatregelen in een verbeterde kwaliteit van preklinische werkstukbeoordelingen? Allerlei deelvraagstellingen zijn af te leiden uit deze centrale vraag. In hoofdstuk VI komen deze deelvragen uitgebreid aan de orde. Hoofdstuk III en IV, waarin respectievelijk de ontwikkeling van de nieuwe beoordelingsmethode en het geïndividualiseerde trainingsprogramma worden besproken, zijn grotendeels bepalend voor de inhoud van de deelvragen.

III DE ONTWIKKELING VAN EEN BEOORDELINGSMETHODE VOOR PREKLINISCHE PRACTICUMWERKSTUKKEN

3.1 Inleiding

Dit hoofdstuk beschrijft de ontwikkeling van een nieuwe beoordelingsmethode voor de klasse II-tweevlakspreparatie voor amalgaam: één van de werkstukken die in het preklinisch motorisch onderwijs vervaardigd moet worden. Bij de constructie van dit nieuwe beoordelingsinstrument is zoveel mogelijk rekening gehouden met resultaten uit onderzoek naar beoordelingsinstrumenten en met de aanbevelingen van de Subcommissie "Toetsing en beoordeling motorische vaardigheden" (zie hoofdstuk II). Het ontwikkelde beoordelingsinstrument wordt gekenmerkt door zeer nauwkeurige omschrijvingen en door het aangebrachte onderscheid in "criterium", "beoordeling" en "scoring". In het vervolg wordt het nieuw ontwikkelde beoordelingsinstrument aangeduid met de term "beoordelingsprotocol".
In dit hoofdstuk wordt achtereenvolgens besproken:
- de argumentatie voor de keuze van de klasse II-tweevlakspreparatie, als preklinisch werkstuk waarvoor een beoordelingsprotocol geconstrueerd is (paragraaf 3.2);
- het beoordelingsprotocol als operationalisatie van de in het preklinisch onderwijs gebruikte beoordelingsmethode (paragraaf 3.3);
- het testen van het beoordelingsprotocol bij docenten en eerstejaars studenten (paragraaf 3.4).

3.2 Waarom een nieuw beoordelingsinstrument?

De beslissing om een nieuwe beoordelingsmethode te ontwikkelen voor gebruik in het preklinisch onderwijs, berust op een drietal argumenten die van onderwijskundige, tandheelkundige en praktische aard zijn.
Het onderwijskundige argument kwam al ter sprake in hoofdstuk I, waar het belang van terugkoppeling, in de vorm van kennis van de resultaten, besproken werd voor de verwerving van vaardigheden. Terugkoppeling kan echter alleen zinvol zijn als de informatie die teruggekoppeld wordt voldoende betrouwbaar is. In par. 2.4.2 is aan de hand van resultaten uit onderzoek aangetoond, dat de momenteel gebruikte beoordelingsmethode voor de klasse II-tweevlakspreparatie geen betrouwbaar instrument is.
Het tandheelkundige argument wordt ontleend aan het werken onder simulatie-omstandigheden. Het oefenen van tandheelkundige vaardigheden op kunststofelementen die in een zogeheten "fantoomkop" zijn geplaatst, heeft als voordeel, dat aan alle studenten een identiek element aangeboden kan worden waarin, bijvoorbeeld, een standaardpreparatie gemaakt moet worden. Deze standaardisatie maakt het mogelijk om prestatiecriteria af te leiden die altijd van toepassing zijn.

Het praktische argument, tenslotte, heeft ook te maken met het simuleren van de patiëntbehandeling. Echte patiënten zullen het naar alle waarschijnlijkheid als uitermate vervelend ervaren, wanneer, na afloop van de behandeling, de instructeur het werk van de student komt beoordelen aan de hand van een zeer uitgebreid beoordelingssysteem. Het twee of meer keren beoordelen van dezelfde prestatie stuit om begrijpelijke redenen op nog grotere bezwaren, zodat controle op de kwaliteit van de beoordeling onmogelijk is.

De keuze voor de klasse II-tweevlakspreparatie als type werkstuk om een nieuwe beoordelingsmethode voor te ontwikkelen werd ingegeven door de volgende overwegingen:
- carieuze aantastingen van het type klasse II komen in de praktijk frequent voor;
- voor het vervaardigen van klasse II werkstukken is een groot deel van alle vaardigheden nodig die bij het prepareren kunnen voorkomen;
- de klasse II-tweevlakspreparatie is een moeilijke preparatie met zeer veel beoordelingsaspecten.

3.3 Twee beoordelingsmethoden voor de klasse II-tweevlakspreparatie

3.3.1 De huidige beoordelingsmethode (kenmerk beoordelingsmethode)

Al vele jaren wordt in het preklinisch motorisch onderwijs van de Nijmeegse Subfaculteit Tandheelkunde de klasse II-tweevlakspreparatie op analytische wijze beoordeeld aan de hand van zes kenmerken. De kenmerken zijn afgeleid uit de principes voor het maken van een caviteitspreparatie, zoals die in de meeste handboeken worden omschreven. Zo is de outline een belangrijk kenmerk vanwege het belang ervan voor de toegankelijkheid, de preventie en de resistentie. Zonder een goede toegankelijkheid wordt het verwijderen van cariës ernstig bemoeilijkt. Om redenen van preventieve aard moeten de begrenzingen van de preparatie in gezond weefsel liggen en in gebieden die minder ontvankelijk zijn voor cariës. De vorm van de caviteit is ook van belang bij het voorkomen van fracturen in element en restauratie (resistentie). De diepte van de preparatie speelt eveneens een rol bij het voorkomen van fracturen in de restauratie. Met de kenmerken caviteit-oppervlaktehoek en pulpo-axiale afschuining wordt beoordeeld of de hoeken, gevormd door de wanden in de caviteit, dié grootte hebben die zowel tandweefsel als restauratie maximaal resistent maken tegen breuken. Convergentie en divergentie van wanden zijn van belang in verband met de retentie van de vulling. Voordat aan de restauratie begonnen wordt moeten alle wanden van de preparatie vrij zijn van oneffenheden en moet al het loszittend tandweefsel verwijderd worden (afwerking).

In de syllabus "prepararen en restaureren" (blok 155/255) (Instituut Conserverende Tandheelkunde voor Volwassenen, 1982) worden voor de zes genoemde kenmerken verschillende prestatiecriteria gespecificeerd. In tabel 3.1 zijn deze opgenomen. Te vaak zijn de criteria in subjectieve termen verwoord (onderstreepte woorden). Het gebruik van een dergelijke beoordelingsmethode heeft tot gevolg dat de inter-beoordelaarsbetrouwbaarheid laag zal zijn. Dit is gebleken uit het in par. 2.4.2 besproken onderzoek van Sanders, die associaties berekende voor 28 beoordelaarsparen en constateerde dat deze matig tot laag waren (Pearson correlatie coëfficiënten varieerden van 0.18 tot 0.65).

Er zijn twee hoofdoorzaken voor de slechte kwaliteit van de beoordelingen aan de hand van de kenmerk-beoordelingsmethode:

1. Het analytische karakter van de kenmerk-beoordelingsmethode is erg oppervlakkig. Er zijn weliswaar 21 prestatiecriteria geformuleerd, maar de uiteindelijke beoordeling van de klasse II werkstukken geschiedt aan de hand van zes kenmerken. Beoordelingsverschillen kunnen zo niet uitblijven. Van beoordelaars kan niet verwacht worden dat ze tot eensluidende beoordelingen komen, wanneer ze niet weten hoe prestaties op de verschillende prestatiecriteria uitgedrukt moeten worden in één kenmerkoordeel. Te veel wordt overgelaten aan de beoordelaars zelf. De verschillen tussen hen met betrekking tot opleiding, ervaring, stemming, enz. veroorzaken interpretatieverschillen die op hun beurt weer leiden tot verschillen in de beoordelingen. Maar, zelfs wanneer er sprake zou zijn van een éénduidige methode om prestaties op de afzonderlijke prestatiecriteria om te zetten in één kenmerk-beoordeling, zouden aanzienlijke verschillen tussen beoordelaars toch blijven bestaan. De gespecificeerde prestatiecriteria zijn namelijk dermate subjectief geformuleerd, dat beoordelen geen constaterende maar een interpreterende activiteit is.
2. Het objectief beoordelen van werkstukken wordt belemmerd door de gebruikte beoordelingsschaal. Elk kenmerk wordt gescoord op een driepunts-schaal:
   1 = de kwaliteit van het kenmerk is slecht en niet meer te verbeteren (onvoldoende);
   2 = het kenmerk voldoet niet geheel aan de eisen, maar is nog wel acceptabel (voldoende);
   3 = het kenmerk is kwalitatief uitstekend (goed).
   De beoordelingsschaal is feitelijk onbruikbaar door het gebruik van subjectieve termen. Een oplossing in de zin van nauwkeurig gedefinieerde schaalpunten is in het kader van de kenmerk-beoordelingsmethode niet mogelijk, omdat op kenmerkniveau objectief geformuleerde prestatiecriteria ontbreken. Deze kunnen ook niet geformuleerd worden, omdat er te veel beoordelingsaspecten aan één kenmerk te onderscheiden zijn.

Tabel 3.1: Criteria voor de klasse II-tweevlakspreparatie.

kenmerk	step	box
outline	1. Al het zwartgemaakte weefsel is weggenomen. 2. De caviteit heeft de juiste breedte. 3. Buccaal en linguaal bestaat de juiste afstand tot de omtrek van het occlusale vlak. 4. De approximale wand waar geen box is gemaakt heeft voldoende dikte.	5. Het geprepareerde element ligt vrij van het buurelement.
diepte	6. De caviteit heeft de juiste diepte. 7. De bodem van de caviteit loopt evenwijdig met het vlak door de knobbeltoppen.	8. De bodem heeft de juiste diepte (1.3 mm). 9. De axiale wand loopt parallel met de buitenkant.
caviteit-oppervlakte-hoek		10. De buccale en linguale wanden staan loodrecht op de buitenkant van het element.
convergentie	11. De buccale en linguale wanden zijn ondersneden. 12. De approximale wand aan de zijde, waar geen box geprepareerd is, is niet ondersneden.	13. De buccale en linguale/palatinale wanden van de box en de step convergeren iets naar occlusaal.
pulpo-axiale afschuining		14. De tweevlakshoek tussen bodem van de step en opstaande wand van de box is afgeschuind (45°) en heeft een breedte van 1/2 mm.
afwerking	15. De outline verloopt vloeiend. 16. De bodem en de opstaande wanden zijn vrij van onregelmatigheden. 17. De preparatie is schoon.	18. De cervicale wand is glad. 19. De outline verloopt vloeiend. 20. De bodem, de axiale wand en de opstaande wanden zijn vrij van onregelmatigheden. 21. De preparatie is schoon.

De conclusie luidt derhalve dat de kenmerk-beoordelingsmethode slechts een oppervlakkige analytische beoordelingsmethode is. De kenmerken zijn te veel omvattend om betrouwbare beoordelingen mogelijk te maken. Analytische beoordelingssystemen vereisen een beoordeling voor elk aspect apart op een beoordelingsschaal met goed gedefinieerde schaalpunten.

3.3.2 Het beoordelingsprotocol (subkenmerk beoordelingsmethode)

3.3.2.1 Inleiding

In 1980 vervaardigde Sanders (1980) een beoordelingsprotocol voor de preparatie en restauratie van vier typen werkstukken uit het preklinisch conserverend onderwijs. Voor elk kenmerk zocht hij in de syllabus "Preparatie en Restauratie blok 155 en 255" naar relevante aspecten en naar de eisen gesteld aan die aspecten. De aspecten van een kenmerk werden subkenmerken genoemd, omdat ze gezamenlijk dat kenmerk operationaliseren. Geheel nieuw was de omschrijving van een beoordelingsmethode voor elk subkenmerk en de daaraan gekoppelde scoringsvoorschriften. In paragraaf 3.3.2.2 wordt uitvoerig ingegaan op de principes van het beoordelingsprotocol.

Het beoordelingsprotocol voor de klasse II-tweevlakspreparatie werd door Straetmans aangepast en verbeterd en in een voorstudie (Straetmans, 1982) uitgeprobeerd bij docenten en studenten. Op grond van de resultaten uit die voorstudie werd het beoordelingsprotocol bijgesteld. Ook de vormgeving werd gewijzigd: het formaat werd teruggebracht van A4 naar ongeveer A5, opdat docenten en studenten het boekje in het borstzakje van hun witte jas zouden kunnen dragen.
Over het doel, de opzet en de resultaten van de voorstudie gaat paragraaf 3.4.

3.3.2.2 Het beoordelingsprotocol

In onderstaande beschrijving van de werkwijze van het beoordelingsprotocol is uitgegaan van de nieuwste versie. Deze kwam tot stand op grond van wijzigingen in een oudere versie, naar aanleiding van resultaten uit een (in par. 3.4 beschreven) pilotstudy. "Beoordelingsprotocol" is een verzamelnaam voor de omschrijving, de beoordelingsmethode en het scoringsvoorschrift van te onderscheiden aspecten aan tandheelkundige werkstukken. De in het preklinisch onderwijs gebruikte kenmerk-beoordelingsmethode is geoperationaliseerd door middel van de subkenmerken in het beoordelingsprotocol. Voor elk subkenmerk wordt omschreven:
1. aan welke eisen werkstukken moeten voldoen (de prestatiecriteria);
2. hoe vastgesteld dient te worden of werkstukken met betrekking

tot het te beoordelen aspect aan de gestelde eisen voldoen;
3. hoe de observatie in een score vertaald moet worden.
Daarnaast is geprobeerd om elk subkenmerk van een illustratie te voorzien om de omschrijving van prestatiecriterium en beoordelingsmethode te ondersteunen.

ad 1. De prestatiecriteria

De prestatiecriteria zijn ontleend aan de in het preklinisch onderwijs gebruikte syllabus voor prepareren en restaureren. In hoeverre de afgeleide subkenmerken werkelijk van invloed zijn op de kwaliteit van het werkstuk is helaas onbekend. Dit validiteitsaspect is nog bijna niet onderzocht, zodat voorlopig de volledige lijst met subkenmerken gehandhaafd blijft.
Inhoudelijk waren er regelmatig verschillen van mening; de definitieve versie van het beoordelingsprotocol is tot stand gekomen op basis van veelvuldig commentaar door tandartsen van het instituut. Uiteindelijk konden zij zich allemaal verenigen met de inhoud.
Aan de formulering van de prestatiecriteria is veel aandacht besteed. Mackenzie's eisen (zie paragraaf 2.3.2) voor prestatiecriteria zijn daarbij gevolgd. Zo is geprobeerd om alle prestatiecriteria in operationele termen te omschrijven en wordt, waar mogelijk, met afmetingen gewerkt. De prestatiecriteria zijn zó omschreven dat duidelijk is wat nog wel en wat niet meer acceptabel is. Bij de meeste subkenmerken wordt dit bewerkstelligd door een onder- en bovengrens van de afmetingen te specificeren (zie figuur 3.1).

ad 2. Het vaststellen of aan de gestelde eisen voldaan wordt

Gestandaardiseerd gebruik van beoordelingsinstrumenten wordt bevorderd door beoordelaars nauwkeurige aanwijzingen te geven over de wijze waarop vastgesteld moet worden of de te beoordelen prestatie aan de gestelde eisen voldoet. In het beoordelingsprotocol zijn voorschriften opgenomen met betrekking tot de beoordelingsmethode, de plaats waar beoordeeld moet worden en de hulpmiddelen die voor het beoordelen gebruikt moeten worden.
Er wordt gebruik gemaakt van drie beoordelingsmethoden: meten, schatten en vergelijken. De breedte en diepte van een preparatie worden gemeten. Hoeken, gevormd door op elkaar staande wanden, worden geschat. De afwerking wordt beoordeeld door die te vergelijken met de afwerking van een referentie-werkstuk.
Waar nodig wordt de plaats waar gemeten, geschat of vergeleken moet worden zeer nauwkeurig aangegeven. Bijvoorbeeld door aan te geven waar precies een instrument ingebracht moet worden. Een voorbeeld uit het beoordelingsprotocol kan dit verduidelijken: "Plaats de rechte sonde in de opening tussen het geprepareerde element en het buurelement. De juiste plaats om in te steken ligt halverwege de afstand tussen het occlusale vlak en de bodem van de box". Vaak verduidelijkt een illustratie de omschrijving.
Bij het beoordelen worden hulpmiddelen gebruikt. Tandheelkundig instrumentarium wordt gebruikt bij de beoordelingsmethoden "meten" en "schatten". In het eerste geval voor echte metingen, in het

geval van "schatten" als hulpmiddel om wanden te verlengen en zo het schatten te vereenvoudigen. Een ander hulpmiddel betreft de referentie-werkstukken. Deze worden gebruikt bij de beoordelingsmethode "vergelijken", als werkstukken beoordeeld moeten worden op afwerking. Het gebruik van referentie-werkstukken is gebaseerd op de overweging dat woorden alleen soms tekort schieten om het verschil tussen "juist acceptabel" en "onacceptabel" aan te geven. Referentie-werkstukken zijn door eerstejaars studenten vervaardigde klasse II- tweevlaks preparaties die nog net voldoende zijn afgewerkt. Het spreekt vanzelf, dat de kwaliteits-aanduiding "net voldoende afgewerkt" van de referentie-werkstukken betrouwbaar moet zijn. Alleen dié werkstukken werden geselecteerd, waarover met betrekking tot de afwerking perfecte overeenstemming bestond tussen zes docenten.

ad 3. Scoring

In paragraaf 2.3.2 werd getwijfeld aan het bestaan van een optimaal aantal schaalpunten voor beoordelingsschalen. De zinvolheid van kwaliteitsonderscheidingen zou bepalend moeten zijn voor het aantal schaalpunten. Om praktische redenen is het verwoorden van een groot aantal absolute kwaliteitsonderscheidingen niet haalbaar; te lange omschrijvingen zouden noodzakelijk zijn. Vandaar dat besloten werd om alleen onderscheid te maken tussen "voldaan" en "niet voldaan aan de eisen". Om het nadeel van de verminderde instructieve waarde van tweepunts-schalen te vermijden, werd voor de meeste subkenmerken de "niet voldaan" categorie gesplitst. Aan de ene "niet voldaan" categorie worden werkstukken toegekend die met betrekking tot het te beoordelen aspect de ondergrens van de vereiste afmeting niet halen, aan de andere "niet voldaan" categorie alle werkstukken die de bovengrens van de afmeting overschrijden.
Voor de meeste subkenmerken wordt dus een nominale beoordelingsschaal gebruikt met drie schaalpunten. De "2"-score betekent dat aan de gestelde eisen voldaan is. Score "1" en "3" houden in dat niet aan de eisen voldaan is en verwijzen elk naar een andere geconstateerde fout.
Voor de subkenmerken waarmee preparaties op hun afwerking worden beoordeeld, wordt een ordinale driepunts-schaal gebruikt. Dit is een consequentie van de gebruikte beoordelingsmethode: het vergelijken met een referentie-werkstuk. De vergelijking kan drie resultaten opleveren: slechter dan het referentie-werkstuk, van gelijke kwaliteit als het referentie-werkstuk en beter dan het referentie-werkstuk. Genoemde resultaten worden respectievelijk gescoord als "1", "2" en "3".

Ter illustratie van de hiervoor beschreven structuur van de subkenmerken wordt in figuur 3.1 een pagina afgebeeld uit de laatste versie van het beoordelingsprotocol. Duidelijk te onderscheiden zijn het prestatiecriterium, de beoordelingsmethode en het scoringsvoorschrift.

1.2. PREPARATIEBREEDTE VAN DE STEP

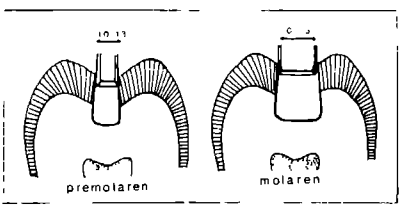
KENMERK OUTLINE

TYPE Klasse II

Bij premolaren mag de breedte van de step varieren van 1.0-1.3 mm.
Bij molaren mag de breedte van de step varieren van 1.0-1.5 mm.

De preparatie is te smal als glazuurmes 10-8-12 niet door de isthmus heen kan en te breed als glazuurmes 13-8-12 en 15-8-12 door de isthmus van resp. premolaar en molaar heen kunnen.

SCORING . Preparatie-breedte is te smal = 1
Preparatie-breedte is juist = 2
Preparatie-breedte is te groot = 3



Figuur 3.1: Een pagina uit het beoordelingsprotocol.

3.4 Pilotstudy naar het functioneren van het beoordelingsprotocol

3.4.1 Doel van de pilotstudy

Eventuele tekortkomingen van het beoordelingsprotocol zouden het snelste aan het licht komen als het in een beoordelingssituatie gebruikt zou worden. Het beoordelingsprotocol werd daarom getest bij een aantal docenten en een aantal eerstejaars studenten. Doel hiervan was het vergaren van informatie over het functioneren van het beoordelingsprotocol. Deze informatie zou op de volgende vragen een antwoord moeten geven:
- zijn de omschrijvingen inhoudelijk juist?
- zijn de omschrijvingen voor iedereen duidelijk?

3.4.2 Materiaal en methode

Docent-beoordelaars

Zes docenten van het Instituut Conserverende Tandheelkunde voor Volwassenen van de Katholieke Universiteit te Nijmegen waren bereid om zes klasse II-tweevlakspreparaties aan de hand van het beoordelingsprotocol te beoordelen en kritische kanttekeningen te plaatsen bij het beoordelingsprotocol. Om inter-beoordelaars-overeenstemmingen te kunnen berekenen vormden de zes docenten drie beoordelaarsparen.

Student-beoordelaars

Van 90 eerstejaars studenten uit het studiejaar 1981-1982 verklaarden zich 32 bereid om mee te doen aan een beoordelingstaak. Ten einde gedegen informatie te krijgen over het beoordelingsprotocol werd besloten om, anders dan bij de docenten, werkstukken op slechts enkele subkenmerken te laten beoordelen. Op deze wijze zouden meer werkstukken aangeboden kunnen worden dan bij de docenten, waardoor de kans op het ontdekken van eventuele tekortkomingen van subkenmerken, vergroot werd. Om slordigheden als gevolg van concentratieverlies te voorkomen, moest het totale aantal te verrichten beoordelingen beperkt blijven tot ongeveer 50. Bij drie subkenmerk-oordelen per werkstuk betekende dit, dat iedere student ongeveer 16 werkstukken zou moeten beoordelen. Om inter-beoordelaarsovereenstemmingen te kunnen berekenen werden zestien beoordelaarsparen gevormd door steeds twee studenten de aangeboden werkstukken op dezelfde subkenmerken te laten beoordelen. De beoordelingen werden verricht in twee sessies met elk 16 studenten en 16 werkstukken. Veelvuldig ruilen van werkstukken was dus onvermijdelijk. Om dat in goede banen te leiden werden in elke sessie vier groepen van elk twee beoordelaarsparen gevormd, waarvoor ruilschema's waren opgesteld. De studenten werden aangemoedigd om onduidelijkheden en tekortkomingen in de omschrijvingen en illustraties, alsmede suggesties voor verbeteringen, schriftelijk te rapporteren. Voor meer gedetailleerde informatie over de opzet van de pilotstudy wordt verwezen naar het eerste voortgangsrapport van het onderwijsstimuleringsproject (Straetmans, 1982).

3.4.3 Resultaten

Informatie over het functioneren van het beoordelingsprotocol werd verkregen door inventarisatie van de geuite kritiek en door berekening van inter-beoordelaarsovereenstemmingen.

3.4.3.1 Kritiek op het beoordelingsprotocol

Docent-beoordelaars

De door de docenten geuite kritiek op het beoordelingsprotocol betrof voornamelijk adviezen voor verduidelijking van omschrijvingen. Inhoudelijk was er nauwelijks kritiek. Over het algemeen waren de docent-beoordelaars van mening dat het beoordelingsprotocol naar bevrediging functioneerde.

Student-beoordelaars

De kritiek die door studenten werd gegeven op het beoordelingsprotocol kan als volgt worden samengevat:

1. Er was twijfel over hoe gescoord moest worden als in hetzelfde werkstuk sprake was van tegengestelde bevindingen met betrekking tot een bepaald beoordelingsaspect.
2. De kwaliteit van de referentiewerkstukken bleek niet ideaal te zijn.
3. Het schatten van hoeken werd als zeer lastig ervaren.
4. Het onderscheid tussen de scoringscategorieën was vaag bij de subkenmerken van het kenmerk "afwerking".

3.4.3.2 Overeenstemming tussen beoordelaars

De overeenstemming tussen beoordelaars is een maat voor de betrouwbaarheid van het beoordelingsinstrument en geeft dus aan in hoeverre men zich kan verlaten op de beoordelingen. Als overeenstemmingsmaten worden het percentage overeenstemming en een afgeleide van coëffiënt Kappa gebruikt. In hoofdstuk V worden deze en nog andere overeenstemmingsmaten uitgebreid besproken.
De berekende overeenstemmingen in deze pilotstudy kunnen niet direct met elkaar vergeleken worden, aangezien beoordelaarsparen òf andere werkstukken beoordeeld hebben (docenten) òf dezelfde werkstukken hebben beoordeeld op andere subkenmerken (studenten).

Docent-beoordelaars

De overeenstemming tussen docenten werd berekend over zes werkstukken en 35 subkenmerken. In tabel 3.2 is voor elk beoordelaarspaar de overeenstemming uitgedrukt in een percentage en in coëfficiënt Kappa. De reële overeenstemming (Kappa) is statistisch getoetst op significant afwijken van nul. De overeenstemmingen zijn matig. Het is mogelijk dat de overeenstemmingen negatief beïnvloed zijn door één of meer groepen van subkenmerken. Door de overeenstemmingen te berekenen voor subkenmerkgroepen apart, kan nagegaan worden of één of meerdere daarvan beduidend lager zijn bij elk beoordelaarspaar. Tabel 3.3 geeft deze overeenstemmingen weer.

Tabel 3.2 : Inter-beoordelaarsovereenstemming docenten.

beoordelaarspaar	overeenstemming %	Kappa
1	65	0.48*
2	57	0.35
3	71	0.56*

* p<.05

Tabel 3.3 : Percentage overeenstemming en reële overeenstemming (Kappa) tussen beoordelaars, per beoordelingsaspect.

subkenmerken van beoord. aspect:		beoordelaarspaar 1	2	3
Outline	%	69	88	84
	Kappa	0.54*	0.81**	0.76**
Diepte	%	66	53	90
	Kappa	0.48*	0.30*	0.85**
Cav.opp.hoek	%	77	70	63
	Kappa	0.65*	0.55*	0.45
Convergentie	%	88	73	71
	Kappa	0.82**	0.59*	0.56*
Pulpo-ax.afsch.	%	83	83	83
	Kappa	0.75**	0.75**	0.75**
Afwerking	%	44	20	55
	Kappa	0.16	-0.20	0.33

* p<.05
** p<.01

Uit tabel 3.3 blijkt dat de docent-beoordelaars de meeste moeite hebben gehad om overeenstemming te bereiken over de afwerking. De vrij lage overeenstemmingen in tabel 3.2 zijn voornamelijk het gevolg van de lage overeenstemmingen op dit beoordelingsaspect.

Student-beoordelaars

De overeenstemmingen, berekend tussen de onervaren eerstejaars student-beoordelaars, zijn vooral van belang vanwege de extra indicatie die ze zijn voor de duidelijkheid van de omschrijvingen (en dus voor de objectiviteit van het beoordelingsinstrument). Tabel 3.5 geeft de overeenstemmingen weer voor de 16 beoordelaarsparen.

Tabel 3.5 : Percentage overeenstemming en reële (Kappa) overeenstemming voor student-beoordelaarsparen.

beoordelaars-paar	overeenstemming %	Kappa	beoordelaars-paar	overeenstemming %	Kappa
1	51	0.27	9	45	0.18
2	58	0.38	10	54	0.31
3	63	0.45*	11	48	0.22
4	57	0.36	12	60	0.41
5	67	0.50*	13	46	0.19
6	48	0.22	14	50	0.25
7	50	0.25	15	56	0.34
8	62	0.42	16	42	0.13

* $p<.05$

Zoals tabel 3.5 laat zien zijn de overeenstemmingen tussen de student-beoordelaars niet hoog. Gemiddeld bedragen het percentage overeenstemming en de reële overeenstemming (Kappa) resp. 54% en 0.31. Slechts twee Kappa coëfficiënten zijn significant groter dan nul op het 5% toetsingsniveau. Ook nu is weer nagegaan of bepaalde subkenmerkgroepen zich in negatieve zin onderscheiden. Tabel 3.6 geeft uitsluitsel daarover.

Tabel 3.6: Gemiddelde overeenstemmingen (% en Kappa) per subkenmerk-groep en standaardafwijkingen (s.d.).

subkenmerkgroep	%	s.d.	Kappa	s.d.
Outline	63	14	0.44	0.21
Diepte	51	15	0.28	0.18
Cav.opp.hoek	53	16	0.30	0.24
Convergentie	53	12	0.30	0.18
Afwerking	49	13	0.23	0.19

Anders dan bij de docent-beoordelaars (zie tabel 3.3) is in tabel 3.6 geen sprake van opvallende verschillen in overeenstemming tussen bepaalde subkenmerk-groepen. De overeenstemmingen zijn over de hele linie vrij laag. Het is mogelijk dat het onervaren zijn van de studenten hier een belangrijke rol heeft gespeeld.

3.4.4 Revisie van het beoordelingsprotocol

De resultaten van de pilotstudy waren aanleiding om veranderingen aan te brengen in het beoordelingsprotocol. De meest ingrijpende verandering betrof het verwijderen van drie subkenmerken van het kenmerk "afwerking". De reden voor verwijdering was gelegen in de kritiek van de docenten op die subkenmerken. Twee subkenmerken bleken dezelfde inhoud te hebben. Het subkenmerk "afronden van interne lijnhoeken" diende volgens de docenten verwijderd te worden omdat hier niet aan voldaan kon worden als gevolg van de vorm van de in het onderwijs gebruikte boren. Het subkenmerk "preparatie moet schoon en droog zijn" was voor werkstukken uit het werkstukkenbestand niet relevant. De nieuwste versie van het beoordelingsprotocol bestaat dus uit 32 subkenmerken (zie bijlage 1).
De zeer lage overeenstemmingen bij de docenten met betrekking tot "afwerking", de kritiek van veel studenten op de vage verschillen tussen de scoringscategorieën "slecht", "acceptabel" en "goed" en op de kwaliteit van de refentie-werkstukken, die niet altijd met "goed" gekwalificeerd konden worden, hebben ertoe geleid dat de beoordelingsmethode voor de subkenmerken van "afwerking" werd herzien. Er kwamen andere referentie-werkstukken, die in plaats van de meest ideale afwerking de nog juist voldoende afwerking representeerden. De definiëring van de schaalpunten van de score-schaal werden in overeenstemming gebracht met de nieuwe beoordelingsmethode. In plaats van te spreken over slechte, acceptabele en goede afwerking werden de schaalpunten gedefinieerd als kwaliteitsvergelijkingen met een referentie-werkstuk (zie 3.3.2.2 onder "scoring").
Minder ingrijpend waren veranderingen naar aanleiding van suggesties om andere instrumenten te gebruiken. Docenten stelden voor om de preparatiebreedte van de step (subkenmerk 1.2) met een glazuurmes te meten in plaats van met een Wesco-stopper. Wesco-stoppers zouden in nauwkeurigheid te veel afwijken ten opzichte van elkaar. Constateren van op- of aflopen van de bodem van de step (subkenmerk 2.3) zou volgens docenten beter gaan met een pocket-sonde dan met een gewone sonde, omdat eerstgenoemde maatstreepjes heeft. Pocketsondes werden door docenten eveneens aanbevolen voor het verlengen van wanden als hulpmiddel bij het schatten van hoeken (subkenmerk 4.1 tot en met 4.6). De door het beoordelingsprotocol voorgeschreven rechte sondes zouden veelal niet recht zijn als gevolg van oneigenlijk gebruik. Alle suggesties werden opgevolgd. Tenslotte werd bij enkele subkenmerken (1.6 en 1.7) de omschrijving van de beoordelingsmethode verduidelijkt door nauwkeuriger aan te geven waar het instrument ingebracht moet worden. Als in het vervolg van dit proefschrift over het beoordelingsprotocol wordt gesproken dan wordt daarmee de gereviseerde, in bijlage 1 gepresenteerde, versie bedoeld.

IV DE ONTWIKKELING VAN EEN GEINDIVIDUALISEERD TRAININGS-PROGRAMMA VOOR BEOORDELAARS

4.1 Inleiding

In paragraaf 2.3.2 zijn enkele onderzoeken besproken waarin beoordelaars* getraind werden in het beoordelen van tandheelkundige werkstukken. Over het algemeen werden geen of slechts geringe trainings-effecten aangetroffen. Niettemin werd toch besloten om een trainingsprogramma te ontwikkelen. De argumenten daarvoor worden hieronder besproken.

1. De besproken onderzoeken naar de effecten van beoordelaarstraining bleken te verschillend van opzet om daaruit algemene conclusies te kunnen trekken over het nut van beoordelaarstraining. In sommige onderzoeken was training een éénmalige activiteit, in andere onderzoeken werden beoordelaars herhaaldelijk getraind. Ook de verschillen in gehanteerde beoordelingssystemen maakten vergelijkingen tussen de trainingsresultaten dubieus.
2. Een ander belangrijk argument om toch een trainingsprogramma te ontwikkelen was de meervoudige gebruiksmogelijkheid van een dergelijk programma. Behalve voor het vergroten van de inter- en intra-beoordelaarsovereenstemming zou een trainingsprogramma gebruikt kunnen worden voor:
   a. zelfevaluatie van studenten;
      Door aan beoordelingstrainingen deel te nemen leren studenten hoe ze hun eigen werkstukken moeten beoordelen. Het is aannemelijk dat leerprocessen positief beïnvloed worden, als studenten in staat zijn om afstand te nemen van hun werkstuk en dit kritisch kunnen bekijken. Daarnaast is het belangrijk dat studenten al tijdens hun opleiding geconfronteerd worden met de praktijksituatie, waarin deskundige beoordeling slechts van henzelf afkomstig kan zijn. In de literatuur zijn enkele onderzoeken te vinden met betrekking tot zelfevaluatie in het tandheelkundig onderwijs.
      Geissler (1973) vond vrij hoge correlaties tussen student- en stafoordelen. Belangrijker waren zijn bevindingen dat, als studenten hun eigen werkstukken beoordeelden, de kwaliteit van de werkstukken verbeterde en minder vaak dezelfde fouten werden gemaakt.
      Abrams en Kelley (1974) rapporteerden dat tweederde van de studenten geen problemen had met zelfevaluatie en dat de meerderheid vond, dat kennis van de criteria van nut was bij het vervaardigen van een preparatie.

*In dit hoofdstuk worden de termen "beoordelaar", "deelnemer" en "trainee" naast elkaar gebruikt als het gaat om personen die aan een beoordelaarstraining deelnemen.

Edwards, Morse en Mitchell (1982) vonden het grootste voordeel van zelfevaluatie, dat het leerproblemen van studenten hielp diagnosticeren. Bijvoorbeeld, als studenten zichzelf hoog waardeerden voor een slechte prestatie, dan zou dit kunnen wijzen op verkeerde interpretatie van de criteria en niet noodzakelijk op problemen met de psychomotorische aspecten van de taak.

b. het inwerken van nieuwe practicum-assistenten;
Beoordelaars hanteren niet altijd dezelfde standaard, komen van verschillende opleidingen, verschillen in ervaring, enz. Dergelijke factoren zijn er de oorzaak van dat instructeurs-wisselingen dikwijls gepaard gaan met wijzigingen in de beoordelingsaanpak. Training kan nieuwe instructeurs snel op de "lijn" van de overige stafleden brengen.

c. het implementeren van nieuwe kwaliteitseisen.
Zoals bij elke tak van wetenschap veranderen ook in de tandheelkunde inzichten en verworvenheden na verloop van tijd. Bijstelling van criteria is dan noodzakelijk. Om verzekerd te zijn van snelle invoering van de nieuwe standaarden kan training verplicht gesteld worden voor alle instructeurs.

3. Het laatste argument om een trainingsprogramma te ontwikkelen werd ingegeven door de overweging, dat in de onderzoeksliteratuur, voor zover bekend, nog niet eerder gewerkt werd met geïndividualiseerde trainingsprogramma's. Het geïndividualiseerde karakter maakt dat het trainen van beoordelaars op minder organisatorische problemen stuit dan bij de gebruikelijke groepstrainingen het geval is. Dit voordeel wordt tezamen met nog enkele andere voordelen besproken in par. 4.2.2.

In de volgende paragrafen van dit hoofdstuk wordt nader ingegaan op de functie van terugkoppeling bij het trainen van beoordelaars (par. 4.2), op twee belangrijke voorzieningen voor geïndividualiseerde trainingen (par. 4.3 en 4.4) en op de opzet van een onderzoek naar het functioneren van het trainingsprogramma en het beoordelingsprotocol (par. 4.5).

4.2 Terugkoppeling: centraal mechanisme in de beoordelaarstraining

4.2.1 Voorwaarden voor terugkoppeling

In de leerpsychologische literatuur wordt een hoge mate van overeenstemming aangetroffen over het belang van invoering en regulering van terugkoppeling van leerresultaten om leerprocessen positief te beïnvloeden. In hoofdstuk I werd de betekenis aangegeven van terugkoppeling voor het aanleren van vaardigheden in het algemeen en tandheelkundige (motorische) vaardigheden in het bijzonder. In hoofdstuk II werd geconstateerd dat in het tandheelkundig onderwijs de kwaliteit van de informatie die teruggekoppeld wordt, veel te wensen overlaat. Naast de constructie van een

kwalitatief goed beoordelingsinstrument bleek het trainen van beoordelaars een veelvuldig toegepaste methode om de kwaliteit van die informatie (de werkstukbeoordelingen) te verbeteren.
Zoals bij elke aan te leren vaardigheid speelt ook bij het leren beoordelen de terugkoppeling van de leerresultaten een belangrijke rol. Buis (1978) onderscheidt twee noodzakelijke voorwaarden die vervuld moeten zijn om van functionerende terugkoppeling te kunnen spreken:

1. "Het scheppen van een situatie, waarin de potentiële informatie-ontvanger zich daadwerkelijk voor terugkoppeling openstelt." Twee vormen van "verstek laten gaan" worden onderscheiden:
   a. De gelegenheid tot het verkrijgen van terugkoppeling wordt niet benut door fysieke absentie.
   b. Men onderwerpt zich aan de terugkoppelingsprocedure, maar de verstrekte informatie wordt niet verwerkt.

   De eerste vorm van "verstek laten gaan" is volgens Buis het gevolg van het ontbreken van duidelijk herkenbare consequenties voor de deelnemers.
   De tweede vorm van "verstek laten gaan" bestaat uit het oneigenlijk gebruik van terugkoppelingsprocedures en uit het afweren van de teruggekoppelde informatie op grond van het egokrenkende karakter ervan.
2. Als aan de eerste voorwaarde voldaan is, moet de informatie op zodanige wijze worden aangeboden, dat de ontvanger er iets mee kan doen.

In de volgende paragraaf wordt onder andere besproken hoe aan deze twee voorwaarden voor terugkoppeling voldaan kan worden in het kader van de beoordelaarstraining.

4.2.2 Terugkoppeling van beoordelingsprestaties

Het beoordelen van een tandheelkundig werkstuk is een perceptueel-cognitieve taak. De combinatie van nauwkeurig waarnemen en gedegen kennis van de criteria waaraan het werkstuk moet voldoen, is een vereiste voor een kwalitatief goede beoordeling. Dat wil zeggen, een beoordeling die representatief is voor de kwaliteit van het werkstuk. In het beoordelingsprotocol wordt veel aandacht besteed aan zowel het waarnemings- als aan het kennisaspect. Desondanks is het beoordelingsprotocol slechts een benadering van een objectief beoordelingsinstrument en kan het dus voorkomen dat sommige subkenmerken verschillend geïnterpreteerd worden door beoordelaars. Verwacht wordt dat de objectiviteit gediend is met het trainen van beoordelaars. Onder training wordt in dit verband verstaan: het herhaald beoordelen van practicumwerkstukken, waarbij de vaardigheid van het beoordelen vergroot wordt als gevolg van de onmiddellijke en gedetailleerde terugkoppeling aan de beoordelaars over hun beoordelingsprestaties. Door de verstrekte terugkoppeling leren de beoordelaars hoe de omschrijvingen in het beoordelingsprotocol moeten worden opgevat en toegepast. Concreter: door de confrontatie met hun eigen prestaties leren de

beoordelaars op dezelfde manier waarnemen, discrimineren, meten, schatten en vergelijken. De training heeft dus tot doel om het onvermijdelijke interpreteren van prestatiecriteria in een bepaalde (voor elke beoordelaar dezelfde) richting te laten gaan. Als gevolg daarvan zal de intra- en inter-beoordelaarsovereenstemming toenemen. Het verstrekken van terugkoppeling over de beoordelingsprestaties impliceert vergelijking van de gegeven beoordelingen met een bepaald criterium, in het trainingsprogramma "referentie-oordeel" genoemd. Op grond van die vergelijking weet de beoordelaar of hij al dan niet juist beoordeeld heeft. Met betrekking tot de referentie-oordelen moeten de volgende vragen gesteld worden:
- In hoeverre kan men zich op de referentie-oordelen verlaten?
- Waaraan worden de referentie-oordelen ontleend?
Als het antwoord op de eerste vraag zou kunnen luiden: "volledig", dan zou dat betekenen dat het ontwikkelen van een beoordelingsprotocol en een trainingsprogramma verspilde energie zouden zijn geweest. Immers, volstrekt betrouwbare beoordelingen impliceren de beschikbaarheid over een volstrekt objectieve beoordelingsmethode. Aangezien zo'n methode (nog) niet bestaat kunnen de referentie-oordelen slechts zó betrouwbaar zijn als de gehanteerde beoordelingsmethode toelaat.
Het antwoord op de tweede vraag is nauw gerelateerd aan het antwoord op de eerste. In de wetenschap dat de huidige beoordelingsmethoden slechts beperkt betrouwbaar zijn, wordt getracht om de betrouwbaarheid te bevorderen door meerdere beoordelaars in te schakelen (zie par. 2.2.2). De referentie-oordelen worden daarom ontleend aan de modale beoordelingen van expert-beoordelaars (zie par. 4.3.3).
Tenzij men kan beschikken over een werkstukkenverzameling waarvan de kwaliteit van de afzonderlijke werkstukken is vastgesteld door een aantal deskundigen, is men in een trainingssituatie voor het bepalen van de referentie-oordelen afhankelijk van de deelnemers aan de training. De referentie-oordelen zijn meer of minder betrouwbaar al naar gelang de overeenstemming tussen de trainees groter of kleiner is.

Met de beantwoording van de hierboven gestelde vragen zijn tevens de bezwaren genoemd van de als groepstraining ingerichte beoordelaarstraining:
1. In verband met de betrouwbaarheid van de referentie-oordelen is het gewenst dat veel beoordelaars aan een training deelnemen. De organisatie daarvan zal dikwijls op moeilijkheden stuiten, met name in het tandheelkundig onderwijs, waar veel parttimers werken die alleen op de zeer drukke practicumuren tegelijkertijd aanwezig zijn.
2. De teruggekoppelde informatie zal niet altijd even overtuigend zijn voor de trainees, aangezien de referentie-oordelen bepaald worden door de beoordelingen van de deelnemers aan de training.
3. De aanwezigheid van collega's kan door sommige beoordelaars als bedreigend worden ervaren, waardoor het niet denkbeeldig is dat zij bewust anders gaan beoordelen dan in het onderwijs.

Om dergelijke bezwaren te vermijden werd besloten om geïndividualiseerde trainingen te organiseren. Bij geïndividualiseerde trainingen zijn deelnemers niet afhankelijk van de aanwezigheid van anderen voor terugkoppeling over hun beoordelingsprestaties. Dit houdt in, dat de referentie-oordelen van tevoren vastgelegd moeten zijn en dus, dat over een werkstukkenbestand beschikt moet worden. Daarnaast moet het trainingsprogramma zodanig geautomatiseerd zijn dat beoordelaars geheel zelfstandig kunnen werken. In het onderhavige geval verstrekt een microcomputer terugkoppeling aan een beoordelaar over zijn beoordelingsprestaties. In par. 4.4 wordt uitvoerig ingegaan op de automatiseringsaspecten van het trainingsprogramma.

In het resterende deel van deze paragraaf wordt nagegaan of het trainingsprogramma voldoet aan de twee door Buis (1978) geformuleerde voorwaarden voor functionerende terugkoppeling (zie par. 4.2.1). Hoewel algemeen van toepassing, zijn de door Buis geformuleerde voorwaarden vooral van kracht in situaties waarin studenten de mogelijkheid wordt geboden om, voorafgaand aan het tentamen, deel te nemen aan één of meer formatieve toetsen, teneinde hiaten in hun kennis tijdig te kunnen opsporen. Met name geldt dit voor de eerste vorm van "verstek laten gaan": "De gelegenheid tot het verkrijgen van terugkoppeling wordt niet benut door fysieke absentie." Voor de beoordelaarstraining in het tandheelkundig onderwijs kan deze vorm van verstek laten gaan geen rol spelen omdat, indien het belang van de training aangetoond kan worden, deelneming verplicht gesteld wordt voor alle docenten uit de doelgroep van de training.
De tweede vorm van "verstek laten gaan" ("Men onderwerpt zich aan de terugkoppelingsprocedure, maar de verstrekte informatie wordt niet verwerkt") is voor de beoordelaarstraining wel belangrijk. Buis noemde twee aspecten van niet-verwerking:

1. oneigenlijk gebruik van de terugkoppelingsprocedure;
2. afweren van de informatie op grond van het ego-krenkende karakter.

ad 1. Het gevaar bestaat, dat beoordelaars de trainingsresultaten gebruiken om hun beoordelingen in het onderwijs te rechtvaardigen. Docenten die mild beoordelen, om daardoor populair te zijn bij de studenten, kunnen zich in de trainingssituatie anders opstellen om te bewijzen dat ze net zo streng beoordelen als andere docenten. De terugkoppeling functioneert in zo'n geval niet, omdat de betreffende docent geenszins van plan is om zijn beoordelingsgedrag in het onderwijs te veranderen. Hoewel minder waarschijnlijk, is een soortgelijke situatie ook denkbaar voor extreem strenge beoordelaars. Genoemde gevaren zullen zich eerder voordoen bij beoordelaarstrainingen in groepsverband dan bij geïndividualiseerde trainingen, omdat bij eerstgenoemde organisatievorm docenten zich sneller bedreigd voelen door de "oplettende" aanwezigheid van collega's.

ad 2. Het is niet ondenkbaar dat beoordelaars de teruggekoppelde informatie niet verwerken omdat die hen in hun eer aantast.

Vooral beoordelaars die overtuigd zijn van de goede kwaliteit van hun beoordelingen, zullen informatie negeren, die niet in overeenstemming is met die opvatting. Belangrijk in dit verband is de wijze waarop terugkoppeling wordt verstrekt. Bij groepstraining is de kans op krenking groot door de aanwezigheid van collega's. Bovendien heeft groepstraining het nadeel dat de referentie-oordelen bepaald worden door de beoordelingen van de deelnemers aan de training. Voor personen die zich snel gekrenkt voelen kan dat een extra reden zijn om zich af te zetten tegen de teruggekoppelde informatie.

Geïndividualiseerde trainingen kennen deze nadelen niet; zowel bij het beoordelen zelf als bij het terugkoppelen van de beoordelingsprestaties zijn de trainees alleen. Een ander voordeel van geïndividualiseerde trainingen is dat de referentie-oordelen "anoniem" zijn en daardoor wellicht meer acceptabel voor beoordelaars die zich bedreigd voelen door de teruggekoppelde informatie.

Zonder dat hiervoor bewijs kan worden aangevoerd lijkt de conclusie gerechtvaardigd, dat met een geïndividualiseerde werkwijze voldaan wordt aan de eerste voorwaarde voor functionerende terugkoppeling, welke luidt: "Het scheppen van een situatie, waarin de potentiële informatie-ontvanger zich daadwerkelijk voor terugkoppeling openstelt."

De tweede voorwaarde luidt dat de informatie op zodanige wijze moet worden aangeboden, dat de ontvanger er iets mee kan doen. Voor de beoordelaarstraining houdt dit in dat de trainees vooral informatie moeten krijgen over hun eigen beoordelingsprestaties. Bij een geïndividualiseerde training wordt beter aan deze voorwaarde voldaan dan bij een groepstraining, waar de beoordelingsprestaties van alle deelnemers besproken moeten worden. De teruggekoppelde informatie zal nog beter kunnen functioneren als de administratie van de training geautomatiseerd is. Deelnemers kunnen dan direct na de training een overzicht krijgen van hun beoordelingsprestaties, waardoor systematische fouten eerder zullen opvallen. De wijze waarop deelnemers aan een geïndividualiseerde training geïnformeerd worden over hun beoordelingsprestaties lijkt te voldoen aan de tweede voorwaarde voor functionerende terugkoppeling.

4.3 Aanleggen van een werkstukkenverzameling

4.3.1 Inleiding

De keuze voor een geïndividualiseerde trainingsopzet maakte het aanleggen van een werkstukkenverzameling noodzakelijk. Van elk werkstuk in de verzameling moet de kwaliteit zijn vastgelegd op de beoordelingsaspecten van de doelvaardigheid (de klasse II-tweevlaks preparatie). De opbouw van het werkstukkenbestand geschiedde in twee fasen:

1. het verzamelen van geschikte werkstukken;
2. het beschrijven van de kwaliteit van de verzamelde werkstukken.

Beide fasen worden achtereenvolgens besproken in de volgende paragrafen.

4.3.2 Verzamelen van geschikte werkstukken

Het belangrijkste doel van de beoordelaarstraining is het bevorderen van de objectiviteit waarmee beoordeeld wordt. Echter, het bereiken daarvan in een trainingssituatie alléén is niet voldoende. Gehoopt wordt, dat positieve transfer zal optreden naar de beoordelingssituatie in het preklinisch onderwijs. Het is dan ook van groot belang dat het beoordelen tijdens de training niet te veel afwijkt van het beoordelen in het onderwijs. Een belangrijk verschilpunt dat niet vermeden kan worden, is het niet aanwezig zijn van de maker van het werkstuk bij de beoordeling in een trainingssituatie. Hoewel in principe gunstig voor de objectiviteit van het beoordelen, beïnvloedt dit verschil de transfermogelijkheden van de trainingsresultaten in negatieve zin. Wil er sprake kunnen zijn van transfer van trainingsresultaten, dan moeten de aangeboden werkstukken tijdens de trainingen in kwaliteit vergelijkbaar zijn met de aangeboden werkstukken in een werkelijke beoordelingssituatie. Doordat sinds enkele jaren werkstukken van studenten worden verzameld, kon vrij eenvoudig aan deze eis worden voldaan.

Nadat het aantal werkstukken dat het bestand zou moeten bevatten op vrij arbitraire wijze bepaald was op 35, werden deze door een docent uit de ruwe verzameling geselecteerd op grond van de volgende aanwijzingen:

1. Selecteer werkstukken per kenmerk en zorg er voor, dat voor elk kenmerk ongeveer even veel werkstukken geselecteerd worden.
2. Per kenmerk moet de helft van het aantal geselecteerde werkstukken niet voldoen aan het criterium van minstens één, van het betreffende kenmerk afgeleid, subkenmerk.
3. De verhouding van molaren en premolaren moet die van het gebit benaderen, dat wil zeggen 3:2.

De selectie-voorschriften zijn gemakkelijker te begrijpen als ze schematisch weergegeven worden. Tabel 4.1 voorziet hierin.

Bij de selectie van de werkstukken werden de aanwijzingen nauwkeurig gevolgd, zodat de structuur van het werkstukkenbestand conform de structuur van het schema is.

4.3.3 Beschrijving van de kwaliteit van de werkstukken

De geselecteerde werkstukken werden met bijbehorend buurelement gepositioneerd in kleine plastic bakjes, waarop identificatienummers waren aangebracht. Om de kans op herkenning van de identificatienummers zo klein mogelijk te laten zijn, werden randomgetallen gebruikt.

Tabel 4.1: Werkstukkenbestand van de klasse II-tweevlaks preparatie (m = molaar; p = premolaar; Ou = outline; Di = diepte; Ca = caviteit-oppervlakte hoek; Co = convergentie; Pa = pulpo axiale afschuining; Af = afwerking).

	selectiecriteria												
	Ou		Di		Ca		Co		Pa		Af		
	m	p	m	p	m	p	m	p	m	p	m	p	tot
De werkstukken voldoen aan alle subkenmerken van het kenmerk waarop geselecteerd wordt.	2	1	2	1	2	1	2	1	1	1	2	1	17
De werkstukken voldoen niet aan minimaal een der subkenmerken van het kenmerk waarop geselecteerd wordt.	2	1	2	2	2	1	2	1	1	1	2	1	18
Totaal	4	2	4	3	4	2	4	2	2	2	4	2	35

Elk werkstuk werd vervolgens beoordeeld door drie, onafhankelijk van elkaar werkende, expert-beoordelaars (beoordelaars met een langdurige ervaring in het beoordelen van preklinische practicumwerkstukken).
De werkstukken werden eerst met de kenmerkmethode beoordeeld en daarna, aan de hand van het beoordelingsprotocol, op subkenmerkniveau. Uit de drie beoordelingen per werkstuk werden referentie-oordelen gedestilleerd, door voor elk beoordelingsaspect op kenmerk- en subkenmerkniveau de modus (meest voorkomende score) te bepalen. In gevallen dat alle drie beoordelaars een andere score hadden toegekend, gaf het oordeel van de auteur van dit proefschrift de doorslag.

4.4 Automatisering van het trainingsprogramma

In het trainingsprogramma zijn de volgende zaken geautomatiseerd:
1. de administratie van de beoordelingsresultaten;
2. de terugkoppeling van de beoordelingsresultaten.

ad 1. De teruggekoppelde informatie bevat naast een vergelijking van het gegeven oordeel met het referentie-oordeel ook een vergelijking van het gegeven oordeel met alle overige, door

andere trainees gegeven, beoordelingen voor hetzelfde werkstuk. Behalve uitvoerig moet de terugkoppeling ook onmiddellijk zijn. Dat wil zeggen dat de terugkoppeling direct na de beoordeling van een werkstuk verstrekt moet kunnen worden. Voor de trainingsleider is het daarom van groot belang dat alle beoordelingen op overzichtelijke wijze bewaard worden. Dit is tevens van belang om de effectiviteit van de training op de beoordelingskwaliteit te kunnen vaststellen. Voor dat laatste doel is namelijk informatie nodig over de beoordelingsprestaties op verschillende tijdstippen in het trainingsverloop. De veelheid van informatie en de snelheid waarmee die informatie op elk willekeurig tijdstip beschikbaar moet zijn, maken administratie door middel van een computer noodzakelijk. Voor het trainingsprogramma wordt gebruik gemaakt van een database programma, dat functioneert op een Exidy^{R} Sorcerer^{R} microcomputer. Het database programma bestaat uit een aantal deelprogramma's, die het mogelijk maken om records (een record is de geadministreerde beoordeling van een werkstuk) in te voeren, te veranderen, te sorteren, te selecteren, uit te breiden, af te drukken (printer) of af te beelden (beeldscherm). Een aantal deelprogramma's hebben een zoekmogelijkheid, waardoor records die niet aan de opgegeven karakteristieken voldoen, kunnen worden overgeslagen. Alle denkbare overzichten van beoordelingsresultaten kunnen daarmee op eenvoudige wijze vervaardigd worden. Zo is het bijvoorbeeld mogelijk om uit het totale recordbestand alleen dié records te selecteren, die de beoordelingen bevatten van beoordelaar X op tijdstip Y. Een van de vele andere mogelijkheden is, dat voor een bepaald werkstuk alle beoordelingen van beoordelaar X, gesorteerd op datum, uit het recordbestand geselecteerd worden. Vooral voor het berekenen van overeenstemmingen binnen en tussen beoordelaars zijn dergelijke overzichten van groot nut.

ad 2. De terugkoppeling van de beoordelingsresultaten is eveneens geautomatiseerd en kent twee vormen:
- terugkoppeling in de vorm van ruwe beoordelingsscores;
- terugkoppeling van beoordelingsresultaten door middel van samenvattende maten, zoals het percentage overeenstemming en coëfficiënt Kappa.

De eerstgenoemde vorm heeft tot doel de trainee te informeren over zijn beoordelingsprestaties door vergelijking met het referentie-oordeel en met de beoordelingen van de andere trainees. Een onderdeel van het hiervoor besproken database programma kan schrapkaarten verwerken waarop trainees de beoordelingen van werkstukken hebben ingevuld. Zodra een schrapkaart is ingelezen gaat het programma op zoek naar het referentie-oordeel van het betreffende werkstuk en naar alle, eventueel door andere trainees gegeven beoordelingen voor dat werkstuk. Een printer drukt direct daarna alle relevante informatie af.

De tweede vorm is vooral van belang om vooruitgang in de beoordelingskwaliteit te kunnen constateren. Bij grote

aantallen beoordelingen is het voor trainees erg moeilijk om vast te stellen of ze het beter hebben gedaan dan de vorige keer. Door de beoordelingsresultaten samen te vatten in één of meer overeenstemmingsmaten wordt de informatie beter toegankelijk. Ook voor de trainingsleider is het uitdrukken van de beoordelingsprestaties in een overeenstemmingsmaat belangrijk. Op deze wijze samengevatte beoordelingsresultaten maken het mogelijk om trainingseffecten vast te stellen en zo het trainingsprogramma te evalueren. Belangrijke eigenschappen van beoordelingssystemen, zoals betrouwbaarheid en validiteit, kunnen eveneens geschat worden via berekening van beoordelaarsovereenstemmingen. In hoofdstuk V wordt uitgebreid hierop ingegaan.
In het trainingsprogramma worden overeenstemmingen berekend tussen beoordelaars en het referentie-oordeel, tussen beoordelaars onderling en binnen beoordelaars. Soms worden de berekeningen uitgevoerd over alle beoordeelde werkstukken, soms over één of enkele. Het komt ook voor dat de overeenstemming berekend moet worden per beoordelingsaspect (kenmerk of subkenmerk) of over een combinatie daarvan.

```
  BEOORDELAARS

Referentie - 1

  KENMERKEN
1 2 3 4 5 6

  WERKSTUKKEN
305 312 364 449 541 644 746 893 897

  FREKWENTIES

   13   4   2
    1  16   1
    0  13   4

  Aantal waarnemingen = 54

OVEREENSTEMMING = .61   Kansovereenstemming = .34
KAPPA = .41
KAPPA-min = -.51 KAPPA-max = .58 KAPPA-relatief = .85
KAPPA RECHTE KANSVERDELING (K_L)= .42
```

Figuur 4.1: Uitvoer-voorbeeld van het programma voor overeenstemmingsberekening.

Door het grote aantal combinaties van beoordelingsscores waarvoor overeenstemmingen berekend kunnen worden, is inschakeling van een computer noodzakelijk. Sanders en Kortsmit (1983) ontwikkelden voor de training een programma voor het berekenen van het percentage overeenstemming en van de reële overeenstemming (Kappa). Het programma vraagt aan de gebruiker tussen welke beoordelaars de overeenstemming berekend moet worden, over welke werkstukken en voor welke beoordelingsaspecten. Tevens is selectie van beoordelingsscores op datum mogelijk. Als de overeenstemming over de opgegeven specificaties berekend is, worden relevante gegevens afgedrukt door een printer. Figuur 4.1 laat dit zien. In dit voorbeeld is de overeenstemming berekend tussen de beoordelingen van beoordelaar 1 en de referentie-oordelen. In totaal zijn negen werkstukken beoordeeld op zes beoordelingsaspecten (kenmerkmethode).

In bijlage 2 is een beschrijving opgenomen van de hierboven besproken software.

4.5 Een onderzoek naar het functioneren van het beoordelingsprotocol en het geïndividualiseerde trainingsprogramma

4.5.1 Inleiding

Het doel van het onderzoek naar het functioneren van het beoordelingsprotocol en het trainingsprogramma is het verkrijgen van een antwoord op de in paragraaf 2.5 geformuleerde centrale vraag, die geconcretiseerd als volgt luidt: Kan de kwaliteit van de preklinische werkstukbeoordelingen verbeterd worden door beoordelaars gebruik te laten maken van het beoordelingsprotocol en door ze te trainen in het beoordelen van practicumwerkstukken? In deze paragraaf wordt de opzet besproken van het onderzoek, dat een antwoord moet geven op deze vraagstelling. Achtereenvolgens komen aan de orde:

- de trainees;
- argumentatie voor het gebruik van twee beoordelingssystemen: de kenmerkmethode en de subkenmerkmethode;
- organisatie in de tijd van het trainingsprogramma;
- de selectie van werkstukken;
- beschrijving van het trainingsverloop.

4.5.2 De trainees

Alle vijf docent-instructeurs van het eerstejaars practicum "Prepareren en Restaureren" (studiejaar 1982-1983) participeerden in het onderzoek. Drie van hen waren ervaren instructeurs; de andere twee waren onervaren. Allen waren bekend met het verschijnsel "beoordelaarstraining" (de coördinator van het betreffende practicum organiseerde zelf op gezette tijden trainingen) en erkenden de noodzaak ervan. Met het onderhavige trainingsprogramma, daarentegen, had geen van hen ervaring.

In een instructie-bijeenkomst werd de werkwijze van het trainingsprogramma uitgelegd en werden afspraken gemaakt voor de eerste trainings-sessie.

4.5.3 Gelijktijdig gebruik van twee beoordelingsmethoden

In het onderzoek werden zowel de kenmerk- als de subkenmerkmethode (beoordelingsprotocol) gebruikt voor het beoordelen van de aangeboden werkstukken. De volgende argumenten werden daarvoor aangevoerd:

1. Het gebruik van beide beoordelingsmethoden zou tijdbesparend kunnen werken. Door de omvang van het beoordelingsprotocol is het beoordelen van werkstukken een langdurig proces. De duur van de training zou aanzienlijk bekort kunnen worden door het beoordelen op subkenmerkniveau afhankelijk te stellen van de beoordeling op kenmerkniveau. Bijvoorbeeld, door in eerste instantie werkstukken alleen op kenmerkniveau te laten beoordelen. Alleen als geen overeenstemming wordt bereikt met het referentie-oordeel moet een trainee voor het betreffende kenmerk de subkenmerken beoordelen. Beoordelaars worden door zo'n opzet extra gemotiveerd om op kenmerkniveau zo nauwkeurig mogelijk te beoordelen. Een nadeel is, dat de hoeveelheid informatie die over de subkenmerken verkregen wordt variabel is. Met name bemoeilijkt dit de kwaliteitscontrole van de referentie-oordelen van de werkstukken uit het werkstukkenbestand.
2. Uitspraken over verschillen in kwaliteit van beoordelingssystemen zouden betrouwbaarder zijn als de gegevens afkomstig zouden zijn van dezelfde werkstukken en dezelfde beoordelaars, onder dezelfde omstandigheden verzameld. Door het beoordelen op subkenmerkniveau afhankelijk te stellen van de beoordeling op kenmerkniveau, kan de vergelijking tussen beide beoordelingsmethoden uiteraard niet optimaal zijn. Echter, verwacht mag worden dat na afloop van de trainingen een voldoende aantal waarnemingen per subkenmerk is verzameld om de vergelijking met de kenmerkmethode te kunnen maken.
3. De kwaliteit van de beoordelingen aan de hand van de kenmerkmethode zou nogmaals onderzocht kunnen worden. Een beslissing over het handhaven of vervangen van de vigerende beoordelingsmethode zou dan genomen kunnen worden op basis van resultaten uit recent onderzoek.

4.5.4 Organisatie van het trainingsprogramma

Voor de organisatie van de training golden de volgende uitgangspunten:

1. Elke beoordelaar moet alle werkstukken uit het bestand tenminste één keer beoordelen. Uitspraken over de kwaliteit van beoordelingssystemen zijn betrouwbaarder als ze gebaseerd zijn op beoordelingen van een groot aantal verschillende werkstukken.

Bovendien is het van belang dat zo veel mogelijk informatie wordt verkregen over de kwaliteit van de werkstukken uit het werkstukkenbestand. Daarmee kan namelijk de kwaliteit van de referentie-oordelen bestudeerd worden.

2. De trainings-sessies moeten elkaar in een snel tempo opvolgen. Grote tussenpozen kunnen de kans op het optreden van trainingseffecten verkleinen.
3. De trainings-sessies mogen niet te lang duren. Vermoeidheid en verveeldheid kunnen snel leiden tot nonchalant beoordelingsgedrag.

Het aantal te beoordelen werkstukken in een trainings-sessie werd arbitrair op minimaal zes bepaald. Bij een werkstukkenbestand van 35 werkstukken betekende dit, dat zes trainings-sessies gepland werden. Elke sessie bestond uit vijf individuele trainingen, die gehouden werden op een voor de trainee en trainingsleider geschikt tijdstip. Na afloop van de laatste individuele training werd een trainings-sessie afgesloten met een plenaire nabespreking. Hierin werden de beoordelingsresultaten van de afgelopen sessie besproken en opvallende zaken ter discussie gesteld. In bijlage 3 wordt een beknopt overzicht gegeven van wat in de plenaire nabesprekingen aan de orde is geweest. Gemiddeld nam een trainingssessie (inclusief plenaire nabespreking) twee weken in beslag. Een individuele training duurde gemiddeld twee uur in de eerste trainings-sessie en 1.5 uur in de laatste trainings-sessie, waarbij aangetekend moet worden dat in de eerste trainings-sessie zes werkstukken werden aangeboden en in de laatste negen.

4.5.5 Werkstukkenselectie

4.5.5.1 Terminologie

In de vorige paragraaf werd beschreven dat in een trainings-sessie minimaal zes werkstukken beoordeeld zouden moeten worden. Het woord "minimaal" kan enige verwarring oproepen, aangezien voor zes trainings-sessies al één werkstuk meer nodig is dan het werkstukkenbestand bevat. De oplossing ligt in het herhaald gebruik van werkstukken. Naast gegevens over de inter-beoordelaarsovereenstemming is het gewenst de beschikking te hebben over intra-beoordelaarsovereenstemmingen. Om die reden werd besloten reeds beoordeelde werkstukken in een der volgende sessies opnieuw aan te bieden (de zogenaamde "Herhaalwerkstukken"). Bovendien werd een werkstuk geselecteerd dat in iedere sessie (steeds onder een ander identificatienummer) aangeboden zou worden. Dit werkstuk werd het "Rode Draad" werkstuk genoemd. In tabel 4.2 wordt schematisch weergegeven hoe de verdeling is geweest van unieke werkstukken, Herhaalwerkstukken en het Rode Draad werkstuk in elke trainingssessie.

Tabel 4.2 : Verdeling van unieke werkstukken, Herhaalwerkstukken en Rode Draad werkstuk in de trainings-sessies.

sessie	aantal unieke werkst.	aantal Herhaalwerkstukken	Rode Draad werkstuk	totaal aantal aangeboden werkstukken
I	6	-	-	6
II	6	-	1	7
III	6	1	1	8
IV	6	1	1	8
V	6	2	1	9
VI	5	3	1	9
totaal	35	7	5	47

In de volgende paragrafen wordt achtereenvolgens de selectie besproken van de unieke werkstukken, van de Herhaalwerkstukken en van het Rode Draad werkstuk.

4.5.5.2 Selectie van de unieke werkstukken

Trainings-effecten zijn eenvoudiger als zodanig te herkennen, als de aangeboden werkstukken gemiddeld zo min mogelijk verschillen van sessie tot sessie. Door gestratificeerde selectie kon dit bereikt worden. De stratificatie van het werkstukkenbestand werd gevormd door het type element (premolaar, molaar) en door de kwaliteit (voldoende of onvoldoende volgens het referentie-oordeel) van de werkstukken. De twee dimensies waarlangs gestratificeerd werd vormden een kruistabel waarin alle werkstukken waren opgenomen.

Tabel 4.3 : Verdeling van de unieke werkstukken naar kwaliteit en type element.

		type element		
		prem.	mol.	Σ
kwaliteit	onvold.	6	15	21
	vold.	7	7	14
	Σ	13	22	35

Gegeven de verdeling van tabel 4.3 en het aantal van zes aan te bieden werkstukken per sessie, werd voor de volgende selectie-verhouding gekozen: 1:3:1:1. Dat wil zeggen dat geprobeerd werd om in elke trainings-sessie 1 onvoldoende premolaar, 3 onvoldoende molaren, 1 voldoende premolaar en 1 voldoende molaar aan te bieden. De werkstukken werden toevallig getrokken uit de stratificaties. In vier van de zes sessies kwam de gewenste verdeling tot stand.

4.5.5.3 Selectie van de Herhaalwerkstukken

De Herhaalwerkstukken werden per toeval getrokken uit de in voorgaande sessies opgenomen werkstukken. Het bestand waaruit getrokken werd bestond uit werkstukken die pas één keer in een sessie waren opgenomen. Het Rode Draad werkstuk en Herhaal-werkstukken uit voorgaande sessies werden dus buiten dit bestand gehouden. Tabel 4.4 kan dit verduidelijken. Volgens tabel 4.2 werd in sessie III voor de eerste keer een Herhaalwerkstuk aangeboden. Het aantal werkstukken waaruit getrokken kon worden bedroeg dus 12 (de unieke werkstukken in de twee voorgaande sessies) minus 1 (het Rode Draad werkstuk in sessie II). Voor de volgende sessies werd op gelijke wijze te werk gegaan.

Tabel 4.4: De omvang van het werkstukkenbestand voor de selectie van Herhaalwerkstukken in sessie III tot en met VI. (U = aantal unieke werkstukken, aangeboden in voorgaande sessies; RD = Rode Draad werkstuk, aangeboden in voorgaande sessies; HH = Herhaalwerkstukken, aangeboden in voorgaande sessies)

sessie	U	- RD	- HH	= selectiebestand
III	12	1		11
IV	18	1	1	16
V	24	1	2	21
VI	30	1	4	25

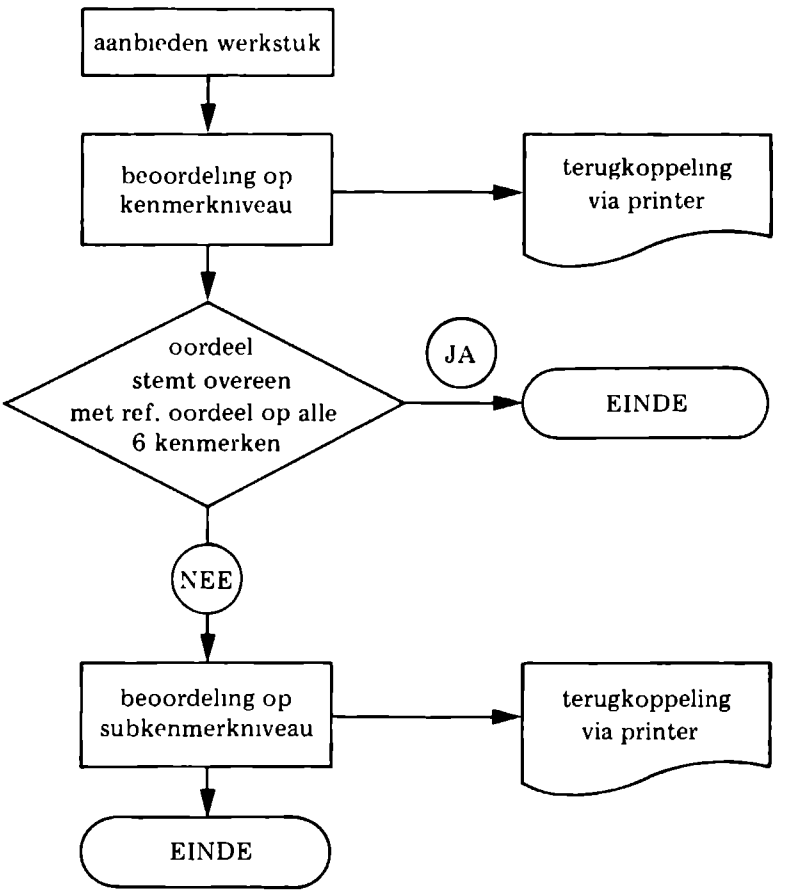
4.5.5.4 Selectie van het Rode Draad werkstuk

Het Rode Draad werkstuk kon pas geselecteerd worden na afloop van de eerste trainings-sessie. Voor de selectie ervan waren namelijk gegevens nodig over de overeenstemming met het referentie-oordeel. Gezocht werd naar een werkstuk waarvoor op kenmerkniveau slechts een lage gemiddelde overeenstemming bereikt werd met het referentie-oordeel. Dit impliceerde namelijk dat dit werkstuk op veel subkenmerken beoordeeld zou moeten worden (als gevolg van de

in par. 4.5.3 besproken koppeling tussen kenmerk- en subkenmerk-beoordelingsmethode), waardoor meer informatie beschikbaar zou komen over de stabiliteit van subkenmerkbeoordelingen.

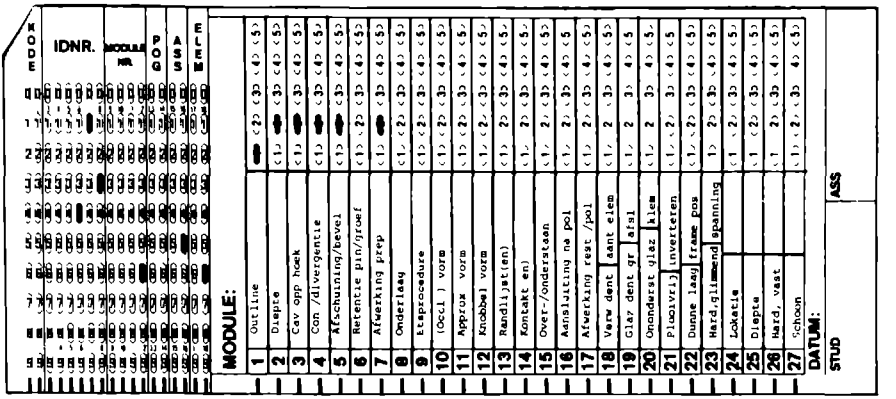
4.5.6 Het verloop van een trainings-sessie

De gang van zaken in een trainings-sessie kan het beste verduidelijkt worden aan de hand van het stroomschema in figuur 4.2.



Figuur 4.2: Stroomschema van het trainings-verloop.

Nadat een werkstuk (bijvoorbeeld werkstuk 413) is aangeboden aan een beoordelaar (bijvoorbeeld beoordelaar 5) gaat deze de kwaliteit ervan beschrijven met behulp van de kenmerkmethode. De bevindingen worden op een schrapkaart ingevuld (zie figuur 4.3), die vervolgens wordt ingelezen.



Figuur 4.3: Schrapkaart voor beoordelingen op kenmerkniveau.

Beoordelaar: 5 Datum: 01-03-83
werkstuk: 413

	kenmerken						
beoorde-laar	Ou	Di	Ca	Co	Pa	Af	Cf
5	1	2	2	2	2	2	6
REF	3	2	2	1	3	2	6
1	1	2	1	1	3	2	4
3	1	3	1	1	3	2	5
2	1	2	2	1	3	2	5
4	1	2	1	1	2	2	6

Beoordeel ook de subkenmerken van: Ou Co Pa

Ou = outline
Di = diepte
Ca = caviteitopper-vlakte hoek
Co = convergentie/ divergentie
Pa = pulpo axiale afschuining
Af = afwerking
Cf = impressionistisch cijfer
REF = referentie-oordeel

Figuur 4.4: Terugkoppeling op kenmerkniveau over de beoordeling van werkstuk 413 door beoordelaar 5.

De computer zoekt het referentie-oordeel op dat behoort bij het opgegeven werkstuknummer en zoekt tevens naar eventuele beoordelingen van andere beoordelaars. Al deze informatie wordt direct na het inlezen van de schrapkaart afgedrukt met behulp van een printer. Figuur 4.4 geeft de uitvoer weer zoals die door de printer wordt vervaardigd, nadat de schrapkaart uit figuur 4.3 is ingelezen.
Omdat het oordeel van beoordelaar 5 op de kenmerken "outline", "convergentie" en "pulpo-axiale-afschuining" niet in overeenstemming is met het referentie-oordeel, geeft de computer opdracht om die beoordelingsaspecten ook op subkenmerkniveau te beoordelen. De beoordelaar raadpleegt nu het beoordelingsprotocol en vult na beoordeling van de betreffende subkenmerken de beoordelingsscores in op een schrapkaart voor subkenmerk-beoordelingen. Na inlezing daarvan wordt wederom onmiddellijke terugkoppeling gegeven over zijn beoordelingsprestaties door middel van printer-uitvoer. Figuur 4.5 laat dit zien.

beoordelaar: 5 werkstuk: 413 datum: 01-03-83

beoordelaar	subkenmerken														
	11	12	13	14	15	16	17	18	41	42	43	44	45	46	51
5	2	2	2	9	9	2	1	2	2	2	3	9	2	2	2
REF	2	2	2	9	9	2	2	2	2	2	3	9	2	2	2
1	2	2	2	9	9	2	1	2							2
3	2	2	2	9	9	2	1	2	2	3	2	9	2	3	
2									2	2	2	9	2	2	2
4	2	2	2	9	9	1	1	1	2	2	2	9	2	2	2

Figuur 4.5: Terugkoppeling op subkenmerkniveau met betrekking tot de beoordeling van werkstuk 413 door beoordelaar 5. De score "9" wordt toegekend als een beoordelingsaspect niet van toepassing is op het type element (molaar/premolaar) dat beoordeeld wordt. Subkenmerk 11 tot en met 18 = outline; subkenmerk 41 tot en met 46 = convergentie; subkenmerk 51 = pulpo ax. afschuining.

De overeenstemming die beoordelaar 5 op subkenmerkniveau bereikt met het referentie-oordeel, is bijna perfect. Alleen voor het

zevende subkenmerk van het beoordelingsaspect "outline" wordt afgeweken van het referentie-oordeel. Nadere inspectie van het terugkoppelingsformulier leert echter, dat ook beoordelaar 1, 3 en 4 geen overeenstemming bereikten met betrekking tot dit subkenmerk. Onderling zijn de beoordelaars het wél eens over de beoordeling van dit aspect. De juistheid van het referentie-oordeel zal geverifieerd moeten worden. De beoordeling van werkstuk 413 door beoordelaar 5 is afgerond en de procedure start opnieuw met een ander werkstuk.

V BETROUWBAARHEID EN VALIDITEIT VAN WERKSTUKBEOORDELINGEN

5.1 Inleiding

Het inschakelen van beoordelaars is een direct gevolg van de onmogelijkheid om de kwaliteit van tandheelkundige werkstukken op volstrekt objectieve wijze (bijvoorbeeld machinaal) vast te stellen.* Als gevolg van het inschakelen van beoordelaars spelen bij de kwaliteitsbepaling van tandheelkundige werkstukken een aantal irrelevante variabelen mee. In de eerste plaats natuurlijk de persoon van de beoordelaar. De combinatie van ervaring, opleiding, karakter, lichamelijk en geestelijk welbevinden maakt van hem een unieke beoordelaar. In de tweede plaats de persoon van de maker van het werkstuk. Los van de ervaringen die de beoordelaar heeft met de maker (zie het vorige punt) kan de aanwezigheid van laatstgenoemde de beoordelaar mild stemmen, onverschillig maken of zelfs agressief. Dezelfde persoon kan totaal andere reacties oproepen bij verschillende beoordelaars. Tenslotte kunnen de omstandigheden tijdens het beoordelen de kwaliteitsbepaling beïnvloeden. Temperatuur, verlichting, beschikbare tijd en rumoer zijn dikwijls plaats- en tijdgebonden en hebben bovendien lang niet altijd dezelfde inwerking op personen. Elke beoordeling van een werkstuk heeft dus een subjectief element in zich, dat de kwaliteit van de beoordeling negatief beïnvloedt. Het is van groot belang (in verband met beslissingen die op grond van beoordelingen genomen moeten worden over studenten en over het onderwijsleerproces) dat over informatie beschikt kan worden met betrekking tot de mate van vertrouwen die in beoordelingen gesteld kan worden. Met andere woorden: hoe betrouwbaar en valide zijn beoordelingen?
In de volgende paragrafen wordt nader in gegaan op het begrip betrouwbaarheid en op de schatters daarvan. In de laatste paragraaf wordt enige aandacht besteed aan de validiteit van werkstukbeoordelingen.

5.2 Betrouwbaarheid van metingen

De betrouwbaarheidsstudie is historisch verbonden met de studie naar individuele verschillen en is grotendeels beperkt gebleven tot gestandaardiseerde intelligentie-, prestatie- en persoonlijkheidstests.

---

*In een poging om werkstukken objectiever te beoordelen construeerden Schiff et al.(1975) een instrument voor het "meten" van de diepte, gladheid en het vlak-zijn van preparaties. Zij rapporteerden significante verbeteringen in de betrouwbaarheid van de metingen.

De klassieke methode van betrouwbaarheidsbepaling van een test houdt in dat voor elk lid van een groep subjecten, waarbij de test wordt afgenomen, twee scores zijn. De mate van consistentie tussen de twee score-reeksen is de betrouwbaarheid van het instrument. Het centrale theoretische concept dat ten grondslag ligt aan deze psychometrische visie op de betrouwbaarheid, is dat elke testscore bestaat uit twee delen:

- een ware score, die de aanwezigheid weergeeft (of de intensiteit) van een of andere trek, eigenschap of gedrag) en
- een foutenscore, die toevallig is en onafhankelijk van de ware score.

Vaak wordt de consistentie uitgedrukt door middel van correlatiecoëfficiënten. Maar andere maten worden ook gebruikt, afhankelijk van het meetniveau van de data en de gewenste interpretatie. Zo kan op het intervalniveau de betrouwbaarheid geschat worden door middel van Pearson's correlatie-coëfficiënt en op ordinaal niveau door Spearman's rangcorrelatie-coëfficiënt. Op nominaal niveau is Pearson's X^2 (chi-kwadraat) een veel gebruikte maat. Sommige maten geven een relatieve, andere een absolute interpretatie van de samenhang tussen score-reeksen. Pearson's correlatie-coëfficiënt is een voorbeeld van een relatieve maat, terwijl de in par. 5.3.2 te bespreken Kappa-coëfficiënt een voorbeeld is van een absolute maat. Het verschil tussen relatieve en absolute samenhang wordt duidelijk aan de hand van het volgende voorbeeld:

		items				
		a	b	c	d	e
beoord.	X	1	2	3	4	5
	Y	6	7	8	9	10

De relatieve samenhang tussen beoordelaar X en Y, uitgedrukt in Pearson's correlatie-coëfficiënt is 1.00. De absolute samenhang, uitgedukt in, bijvoorbeeld, het percentage identieke beoordelingen, is 0. Anders gezegd: Pearson's correlatie coëfficiënt drukt de associatie tussen X en Y uit en het percentage identieke beoordelingen is een uitdrukking van de overeenstemming tussen X en Y.

De interpretatie van een betrouwbaarheidsmeting is ook afhankelijk van de procedure die gevolgd wordt om de betrouwbaarheid te bepalen. Drenth (1973) noemt twee hoofdwegen om betrouwbaarheid te schatten. Ten eerste de weg van herhaald testonderzoek en ten tweede een bepaalde bewerking van één enkel testonderzoek. Bij herhaald testonderzoek wordt dezelfde test twee keer afgenomen (test-hertestmethode) of wordt de tweede keer een equivalente test afgenomen (parallelvorm methode). Binnen de methode van een bewerking van de testresultaten uit één enkel onderzoek bestaan ook twee verschillende mogelijkheden. Ten eerste dié, waarbij men de totale groep opgaven in twee gelijke helften splitst, waaruit

afzonderlijke scores berekend worden. Het verband tussen deze scores geeft een schatting van de betrouwbaarheid. Dit is de split-half methode. Ten tweede is er een methode die gebaseerd is op een variantie-analyse van de antwoorden op de items. Deze methode gaat uit van de interrelatie tussen de testitems afzonderlijk.
Onafhankelijk van de gevolgde procedure krijgt de variantie die toegeschreven wordt aan de individuele verschillen (de ware variantie) gewoonlijk steeds dezelfde interpretatie. Ze geeft stabiele verschillen weer tussen individuen: het ware score deel van de data. De variantie die toegeschreven wordt aan de meetfouten, echter, staat bloot aan variërende interpretaties, afhankelijk van hoe de twee scores verkregen werden. De "fout" omvat uiteraard altijd de echte fout: dié toevalsfluctuaties (zoals bijvoorbeeld gezondheid, mentale toestand, verlichting, temperatuur, enz.) van de talloze factoren die het gemetene kunnen beïnvloeden. Maar de "fout" omvat ook variatiebronnen die afhankelijk zijn van de gehanteerde methode om twee scores te verkrijgen.

Tot zover is alleen in testtheoretische termen gesproken over de betrouwbaarheid van metingen. Beoordelingen, echter, vormen een aparte klasse van psychologische metingen waarvoor de besproken procedures veelal niet geschikt zijn. In het onderhavige onderzoek gaat het bij betrouwbaarheid vooral om de mate waarin, vanuit een bepaalde beoordeling, gegeneraliseerd kan worden naar andere beoordelaars. Deze scorings-betrouwbaarheid kan met behulp van de volgende procedures geschat worden:
- Werkstukken worden door twee of meer beoordelaars gescoord met hetzelfde instrument (kenmerk- of subkenmerkmethode). De inter-beoordelaarsbetrouwbaarheid wordt dan gegeven door de samenhang tussen de score-reeksen.
- Werkstukken worden twee of meer keer door dezelfde beoordelaar gescoord met hetzelfde instrument. De resulterende samenhang is dan een maat voor de intra-beoordelaarsbetrouwbaarheid.

In de eerste procedure geeft de ware score echte verschillen weer tussen werkstukken, terwijl de foutenscore verschillen tussen de beoordelaars weergeeft in het gebruik van het beoordelingsinstrument én toevallige verschillen.
In de tweede procedure weerspiegelt de ware score weer echte verschillen tussen werkstukken en de foutenscore, naast toevallige verschillen, inconsistent gebruik van het beoordelingsinstrument door de beoordelaar.

Zoals al eerder werd opgemerkt, kan bij de betrouwbaarheid van beoordelingen onderscheid gemaakt worden tussen associatie en overeenstemming. De verschillen tussen associatie en overeenstemming zijn gebaseerd op de definities van hun indices. De associatie tussen beoordelaars wordt gewoonlijk uitgedrukt in een of andere correlatie coëfficiënt, die de variantie van een reeks scores verdeelt in ware variantie (verschillen tussen werkstukken) en foutenvariantie. De foutenvariantie kan bestaan uit toevallige fluctuaties in de prestatie van de student, inconsistenties in het gebruik van het beoordelingsinstrument, verschil-

len tussen beoordelaars, enz. De overeenstemming tussen beoordelaars wordt meestal uitgedrukt in een percentage en bevat geen informatie over verschillen tussen werkstukken, maar alleen over één mogelijke foutenbron: verschillen tussen beoordelaars. Anders gezegd: een associatie-coëfficiënt weerspiegelt de relatieve omvang van de foutenscore ten opzichte van de ware score, terwijl de overeenstemming de absolute omvang van één soort fout weergeeft.
Light (1973) illustreert het verschil tussen associatie en overeenstemming aan de hand van een voorbeeld met nominale beoordelingscategorieën. In beide kruistabellen van figuur 5.1 hebben twee leerkrachten onafhankelijk van elkaar ieder 30 studenten toegewezen aan één van de drie gedragscategorieën (A, B of C).

(a)

leerkracht 1 \ leerkracht 2	A	B	C
A		10	
B			10
C	10		

perfecte associatie zonder dat sprake is van overeenstemming

(b)

leerkracht 1 \ leerkracht 2	A	B	C
A	10		
B		10	
C			10

perfecte associatie met perfecte overeenstemming

Figuur 5.1: Illustratie van het verschil tussen associatie en overeenstemming.

In kruistabel a geeft kennis van de categorie waaraan leerkracht 1 iedere student heeft toegewezen perfecte informatie over de toewijzing door leerkracht 2. Er is dus sprake van een perfecte associatie. Maar tussen de twee leerkrachten is geen enkele overeenstemming over de gedragskenmerken van de 30 studenten. In kruistabel b is zowel sprake van perfecte associatie als van perfecte overeenstemming. De overeenstemming weerspiegelt de mate waarin beoordelaars een bepaald subject exact aan dezelfde categorie toewijzen.

5.3 Schatters van de intra- en inter-beoordelaarsbetrouwbaarheid

5.3.1 Inleiding

Voor het schatten van de intra- en interbeoordelaarsbetrouwbaarheid bestaan verschillende maten, toepasbaar bij gebruik van data van nominaal, ordinaal en interval niveau.
Nominale data zijn afkomstig van classificaties zonder rangorde. De beoordelingsschalen die voor de subkenmerken (behalve de

subkenmerken met betrekking tot "afwerking") van het beoordelingsprotocol worden gebruikt, zijn van dit meetniveau (zie par. 3.3.2.2 onder punt 3).
Ordinale data impliceren rangordening van prestaties op een of ander continuum, evenwel zonder dat er sprake is van gelijke intervallen tussen de "rangen". De beoordelingsschaal van de kenmerkmethode is een goed voorbeeld (zie par. 3.3.1).
Data van intervalniveau maken het mogelijk dat uitspraken worden gedaan over de omvang van verschillen tussen scores. Een mogelijke uitspraak is bijvoorbeeld: het verschil in score tussen X en Y is twee keer zo groot als het verschil in score tussen Y en Z. Maar op grond van een dergelijke constatering mag niet geconcludeerd worden dat het verschil in prestatie tussen X en Y twee maal zo groot is als het verschil in prestatie tussen Y en Z. Alleen het verschil in testscores is twee maal zo groot.
Metingen in de gedragswetenschappen zijn zelden van intervalniveau, maar vaak wordt uitgegaan van de assumptie dat een metrische schaal benaderd wordt (Guilford, 1973). In onderhavig onderzoek wordt eveneens uitgegaan van de assumptie dat aan practicumwerkstukken toegekende cijfers van intervalniveau zijn. Deze cijfers komen tot stand via een transformatie van gesommeerde kenmerkscores en liggen op een tienpunts-schaal.
Uit bovenstaande blijkt dat het onderzoek naar het functioneren van het beoordelingsprotocol en het trainingsprogramma metingen oplevert van nominaal, ordinaal en intervalniveau. Omdat de beoordelingsresultaten van de kenmerkmethode (ordinale schaal) vergeleken moeten worden met die van de subkenmerkmethode (nominale schaal), verdient het aanbeveling om de resultaten van beide methodes met dezelfde statistische maat te beschrijven. Gelet op het laagste meetniveau impliceert dit een keuze voor een nominale maat. Cijfers toegekend aan werkstukken worden geanalyseerd met een interval maat.

In par. 5.2 werd aangetoond dat het zinvol is om de betrouwbaarheid te schatten via berekening van zowel de associatie als de overeenstemming tussen beoordelaars. Literatuuronderzoek (Kendall en Stuart, 1961; Landis en Koch, 1975; Everitt, 1977) wees uit dat voor nominale data Tschrupow's T (een functie van Pearson's X^2) de beste associatiemaat was. De andere twee in aanmerking komende maten vielen af omdat in het ene geval de waarde afhankelijk bleek te zijn van de steekproefgrootte (Pearson's X^2) en in het andere geval de maximale waarde niet of bijna niet bereikt kon worden (contingentie-coëfficiënt P).
Ondanks het feit dat een geschikte associatiemaat beschikbaar is voor data van nominaal meetniveau wordt deze niet gebruikt voor het beschrijven van de beoordelingsresultaten in onderhavig onderzoek. De reden hiervoor is gelegen in het feit dat Tschrupow's T lastig te interpreteren is wanneer men geen kennis heeft van de beoordelingsscores waarover T berekend is. Alleen met de bewuste kruistabel voor ogen kan de betekenis van een bepaalde T-waarde goed ingeschat worden. De voorbeelden in figuur 5.2 zijn een illustratie voor de interpretatie-problemen van Tschrupow's T.

(a)

X \ Y	1	2	3
1	8	3	0
2	1	41	2
3	2	2	0

%=83; T=0.52

(b)

X \ Y	1	2	3
1	15	3	0
2	1	20	2
3	2	2	14

%=83; T=0.75

(c)

X \ Y	1	2	3
1	15	0	10
2	0	20	0
3	0	0	14

%=83; T=0.82

Figuur 5.2: Overeenstemming (%) en associatie (T) tussen beoordelaar X en Y in drie verschillende situaties.

In kruistabel a is 75 procent van de beoordelingsscores van X een "2". Voor Y is dit percentage 78. Reynolds (1977) toonde aan dat beperkte variatie in de scores de relatie in de kruistabel zou verzwakken. In kruistabel a wordt dit geïllustreerd. In kruistabel b is de variatie in de beoordelingen van X en Y toegenomen. Terwijl de overeenstemming in deze kruistabel niet veranderd is, is de associatie met meer dan 44 procent toegenomen. In kruistabel c is de overeenstemming nog steeds ongewijzigd terwijl de associatie nog verder is toegenomen tot 0.82. Maar dat een hoge T-waarde niet per definitie gunstig is wordt duidelijk als kruistabel c goed bestudeerd wordt. Behalve de cel die gevormd wordt door het "1"-oordeel van X en het "3"-oordeel van Y zijn alle niet-diagonaalcellen leeg. Bij niet-overeenstemming betreft het dus steeds een situatie waarin X een "1" geeft en Y een "3". Deze systematiek roept twijfels op over de objectiviteit van de gebruikte beoordelingsmethode. Uit dit voorbeeld wordt duidelijk dat het weinig zinvol is om naast een overeenstemmingscoëfficiënt ook nog Tschrupow's T te geven, als niet tevens de kruistabellen gepresenteerd kunnen worden waarover beide maten berekend zijn. Door plaatsgebrek is dit helaas onmogelijk en wordt volstaan met het presenteren van overeenstemmings-coëfficiënten.

In de volgende paragrafen worden de statistische maten besproken waarmee in hoofdstuk VI de intra- en inter-beoordelaarsbetrouwbaarheid geschat zullen worden. Achtereenvolgens zijn dat:
- coëfficiënt Kappa, een overeenstemmingsmaat voor nominale data (par. 5.3.2);
- de intraklasse correlatie coëfficiënt, een associatiemaat op interval niveau (par. 5.3.3);
- index T, een overeenstemmingsmaat voor interval data (par. 5.3.4).

5.3.2 Coëfficiënt Kappa

Een veel toegepaste overeenstemmingsmaat is de proportie of het percentage overeenstemming. De belangrijkste eigenschappen van deze maat zijn de eenvoudige interpreteerbaarheid en het gemak waarmee hij berekend kan worden. Een belangrijk nadeel is dat het percentage overeenstemming een overschatting is van de werkelijke overeenstemming omdat deze maat geen rekening houdt met het feit dat enige overeenstemming verwacht kan worden op basis van kans alleen (Mitchell, 1979). Een maat die het probleem van de kansovereenstemming oplost is coëfficiënt Kappa (Cohen, 1960). Cohen definieerde een gestandaardiseerde coëfficiënt voor overeenstemming op nominaal niveau, die er in formulevorm als volgt uitziet:

$$\kappa = \frac{P_o - P_c}{1 - P_c}$$

Voor een goed begrip van de formule is enige kennis nodig met betrekking tot notatie voor proporties in contingentie-tabellen. Tabel 5.1 is illustratief voor de wijze van noteren. De notatie "P21", bijvoorbeeld, staat voor de proportie werkstukken waaraan door beoordelaar 1 een "2" is toegekend en door beoordelaar 2 een "1".

Tabel 5.1 : Notatie voor proporties in een contingentie-tabel

		beoord. 2				
		1	2	...	L	Σ
beoord. 1	1	P11	P12	...	P1L	P1.
	2	P21	P22		P2L	P2.
	.	.	.		.	.
	.	.	.		.	.
	L	PL1	PL2	...	PLL	PL.
	Σ	P.1	P.2	...	P.L	1

De proportie waargenomen overeenstemming wordt gegeven door de volgende formule:

$$P_o = \sum_{i=1}^{L} P_{ii} \quad (L = \text{aantal schaalpunten})$$

Bij volledige onafhankelijkheid van de beoordelaars wordt de proportie verwachte overeenstemming geschat door:

$$P_c = \sum_{i=1}^{L} P_{i.} P_{.i}$$

Kappa varieert theoretisch tussen $-P_c/(1-P_c) \leq \kappa \leq 1$ en heeft een waarde van 0 als $P_o = P_c$. Een Kappa coëfficiënt wordt getoetst op significant afwijken van 0 door de gevonden waarde te delen door de standaardfout SE(K). Toetsing vindt plaats tegen de normale verdeling door middel van z-waarden: Z = K/SE(K). De standaardfout wordt bepaald met de volgende formule (Fleiss, Cohen & Everitt, 1969):

$$SE(K) = \sqrt{\frac{1}{N(1-P_c)^2}\left(P_c + P_c^2 - \sum_{i=1}^{L} P_{i.}P_{.i}(P_{i.} + P_{.i})\right)}$$

Coëfficiënt K, zoals gedefinieerd door Cohen, zal vrij vaak lage waarden opleveren en kan voor bepaalde toepassingen een conservatieve overeenstemmingsmaat genoemd worden. Dat is een direct gevolg van de definitie van de kansovereenstemming:

$$P_c = \sum P_{i.}P_{.i}$$

De term Pc is afhankelijk van de marginale proporties, waardoor ook in situaties dat aan beoordelaars bekend is welke marginale waarden ze moeten reproduceren, een goede schatting van de reële overeenstemming gewaarborgd is. Maar, in het concrete geval van het beoordelen van tandheelkundige practicumwerkstukken is geen sprake van vastgelegde marginalen die aan de beoordelaars bekend zouden zijn. Brennan en Prediger (1981) zeggen hierover: "Wanneer beoordelaars geen marginale restricties opgelegd hebben gekregen, is elke overeenstemming tussen beoordelaars in de marginale proporties "echte" overeenstemming, of tenminste even reëel als de overeenstemming die aanwezig is op de diagonaal. Voor Cohen's Kappa neemt de index voor kansovereenstemming, $\sum P_{i.}P_{.i}$, toe met een toename in de marginale overeenstemming. Daardoor moeten twee beoordelaars die onafhankelijk en zonder voorafgaande kennis tot een gelijke marginale verdeling komen, een veel hogere overeenstemming bereiken om op een zekere waarde van Kappa te komen, dan twee beoordelaars die totaal verschillende marginalen produceren." Figuur 5.3 illustreert dit heel duidelijk.

	1	2	3	Σ
1	0	1	0	1
2	1	8	1	10
3	0	1	0	1
Σ	1	10	1	12

$P_o = 0.67$
$P_c = 0.71$
$\kappa = -0.14$

	1	2	3	Σ
1	4	0	0	4
2	0	4	0	4
3	0	4	0	4
Σ	4	8	0	12

$P_o = 0.67$
$P_c = 0.33$
$\kappa = 0.50$

figuur 5.3: Cohen's Kappa berekend over een kruistabel met gelijke en verschillende marginale totalen.

Cohen's Kappa beloont beoordelaars dus niet voor het produceren van overeenstemmende marginalen. Het is daarom niet redelijk om, als de marginalen vrij zijn (dat wil zeggen niet op voorhand bekend aan de beoordelaar), de index voor kansovereenstemming te laten afhangen van de geproduceerde marginalen. Bij vrije marginalen kan iedere beoordelaar elke willekeurige set van marginalen produceren, met dié beperking dat er N werkstukken te classificeren zijn in L categorieën. In het geval dat twee beoordelaars op volstrekt willekeurige wijze werkstukken toekennen aan categorieën, is de verwachte marginale proportie voor elke categorie 1/L voor beide beoordelaars. Dus, vóórdat feitelijk wordt toegewezen is de waarschijnlijkheid dat twee beoordelaars, op basis van kans, een werkstuk toekennen aan dezelfde categorie $(1/L)(1/L) = 1/L^2$. Voor alle categorieën is de som van de waarschijnlijkheden gelijk aan $L/L^2 = 1/L$. Als beide marginalen vrij zijn wordt de kansovereenstemming dus gegeven door 1/L. Voor de situatie waarin één marginaal vrij is laten Brennan en Prediger (1981) zien dat ook dan 1/L de beste index is voor kansovereenstemming. De andere definitie van de verwachte overeenstemming resulteert in de volgende formule voor een overeenstemmingscoëfficiënt die corrigeert voor kansovereenstemming:

$$\kappa_L = \frac{\sum P_{ii} - 1/L}{1 - 1/L}$$

Deze formule is een in gewijzigde vorm overgenomen formule voor een coëfficiënt van consistentie, die al in 1954 door Bennett et al. (1954) werd voorgesteld. Als in het vervolg van dit proefschrift over Kappa gesproken wordt, dan wordt κ_L bedoeld. Als over Cohen's Kappa gesproken wordt, dan gaat het over een overeenstemmingscoëfficiënt waarvan de kansovereenstemming bepaald wordt door de marginale proporties. Voor het berekenen van de overeenstemming tussen beoordelaars over de beoordelingsscores afkomstig uit de trainings-sessies, zal gebruik gemaakt worden van κ_L.

Scott (1955) noemt als nadelen van deze formule dat de overeenstemming kunstmatig groot kan zijn als niet-functionele categorieën gebruikt worden en dat de index gebaseerd is op de veronderstelling, dat alle categorieën in de schaal een gelijke waarschijnlijkheid hebben om gekozen te worden. Lawlis & Lu (1972) erkennen Scott's bezwaren maar zijn niettemin van mening dat 1/L goed bruikbaar is als een ondergrens voor de onbekende waarschijnlijkheid van kansovereenstemming.
Een ander nadeel is dat in de literatuur geen passende variantieformule voor κ_L beschreven is. Ten einde de κ_L- coëfficiënten toch te kunnen toetsen ontwikkelde Sanders (Straetmans en Sanders, 1984) een indirecte methode voor het toetsen van κ_L. Een voorbeeld kan de methode illustreren. Twee beoordelaars hebben een werkstuk op tien aspecten beoordeeld op een driepunts-schaal. Hoe groot moet κ_L zijn om op het vijf procent toetsingsniveau significant te zijn? Eerst wordt voor elk mogelijk aantal overeenstemmingen (0-10) berekend op hoeveel manieren dit bereikt kan

worden (zie tabel 5.2). Uit de tabel kan opgemaakt worden dat er drie procent kans is dat twee beoordelaars overeenstemming bereiken over acht of meer beoordelingen. Berekend kan worden dat een overeenstemmingsproportie van 0.758 vereist is voor significantie op het vijf procent toetsingsniveau. Omdat κ_L een lineaire transformatie is van het percentage overeenstemming ($\kappa_L = 0.5\ (3p_o - 1)$, waarbij p_o = proportie waargenomen overeenstemming), kan berekend worden dat de kritische κ_L-coëfficiënt in dit concrete geval gelijk is aan 0.637. Kappa's groter of gelijk aan 0.637 zijn significant op het vijf procent niveau.

Tabel 5.2: Kansverdeling voor overeenstemming bij 10 beoordelingen.

overeenstemming absoluut	relatief	combinaties	kans (%)	cumulatieve kans (%)
0	0	3003	7	100
1	0.1	6006	14	93
2	0.2	7722	17	79
3	0.3	7920	18	62
4	0.4	6930	16	44
5	0.5	5292	12	28
6	0.6	3528	8	16
7	0.7	2016	5	8
8	0.8	945	2	3
9	0.9	330	1	1
10	1	66	0	0
totaal		43758		

5.3.3 Intraklasse correlatie coëfficiënt

Een veel gebruikte maat voor de inter-beoordelaarsbetrouwbaarheid van interval data is de betrouwbaarheids-coëfficiënt r_{kk}. Dit is de bekende Pearson's product moment correlatie coëfficiënt, berekend over de beoordelingsscores van twee beoordelaars. Deze coëfficiënt geeft de proportie weer van de totale variantie die niet aan foutenvariantie te wijten is en tevens de lineaire associatie tussen toegekende cijfers van twee beoordelaars. De gekwadrateerde correlatie-coëfficiënt geeft de proportie verklaarde variantie weer; dat wil zeggen de proportie variantie in de cijfers van de ene beoordelaar die voorspelbaar is vanuit de kennis van de cijfers van de andere beoordelaar. Coëfficiënt r_{kk} varieert van -1.00 tot +1.00. Een r_{kk} van 0 geeft aan dat er geen associatie is tussen de cijfer-reeksen van twee beoordelaars. Een

r_{kk} van 1.00 duidt op een perfecte associatie tussen de beoordelaars (de cijfers van de ene beoordelaar zijn perfect te voorspellen vanuit de kennis van de cijfers van de andere beoordelaar) en ook op een perfecte overeenstemming in de zin van identieke standaardscores. Een r_{kk} van -1.00 duidt op een perfect negatief verband tussen beoordelaars. Een hoog cijfer bij de ene beoordelaar impliceert een laag cijfer bij de andere beoordelaar en andersom (Hartmann, 1977).

Een andere, betere maat voor de associatie tussen beoordelaars voor data op interval-niveau is de intraklasse correlatie-coëfficiënt (symbool R). In de eerste plaats omdat R berekend kan worden uit beoordelingen afkomstig van meer dan twee beoordelaars. In de tweede plaats omdat de intraklasse correlatie-coëfficiënt de mogelijkheid biedt om de op te nemen variantie-componenten te specificeren. "R" wordt geschat op basis van variantie-schattingen afkomstig uit een variantie-analyse en kan geïnterpreteerd worden als de proportie van de totale variantie in de beoordelingen, die toegeschreven kan worden aan de variantie in de werkstukkenkwaliteit. Waarden die de bovengrens van R (1.00) benaderen geven een hoge associatie aan tussen de variantie in de werkstukkenkwaliteit en de totale variantie en wijzen dus op een hoge betrouwbaarheid van de beoordelingen. Een R van 0 geeft aan dat van betrouwbare beoordelingen geen sprake is. Negatieve R-waarden worden, hoewel mathematisch mogelijk, zelden waargenomen (Tinsley en Weiss, 1975).

Er zijn verschillende versies van de intraklasse correlatie-coëfficiënt (ICC), die zeer uiteenlopende resultaten kunnen geven als ze op dezelfde data worden toegepast. Elke versie is geschikt voor de specifieke situatie die gedefinieerd wordt door de onderzoeksopzet en het onderzoeksdoel. Voor een interbeoordelaars betrouwbaarheidsstudie, waarin elk werkstuk uit een toevallig getrokken steekproef (n) beoordeeld wordt door k beoordelaars, bestaan drie verschillende gevallen:

1. Elk werkstuk wordt beoordeeld door een andere groep van k beoordelaars, die puur toevallig geselecteerd werden uit een grotere populatie van beoordelaars.
2. Een toevallige steekproef van k beoordelaars wordt geselecteerd uit een populatie en elke beoordelaar beoordeelt elk werkstuk. Dat wil zeggen: elke beoordelaar beoordeelt n werkstukken.
3. Elk werkstuk wordt beoordeeld door elk van dezelfde k beoordelaars, die de enige beoordelaars zijn waarnaar de interesse uitgaat.

Elk van deze gevallen vereist een verschillend gespecificeerd statistisch model om de resultaten mee te beschrijven. In geval 1 zijn het beoordelaars-effect, het interactie-effect tussen beoordelaar en werkstuk en de toevallige fout niet te scheiden. Deze drie effecten zijn opgenomen in één residu-component. De modellen voor geval 2 en 3 nemen genoemde effecten wel elk afzonderlijk op. Geval 2 verschilt onder andere van geval 3 in de assumpties met betrekking tot het beoordelaars-effect. In geval 2 is het beoor-

delaars-effect een random variabele terwijl in geval 3 het beoordelaars-effect "fixed" is.
In het concrete geval van de beoordelaarstraining is geval 3 van toepassing. Vijf docenten uit het preklinisch onderwijs beoordelen elk alle werkstukken uit het werkstukkenbestand. Aangezien deze vijf docenten geen toevallige steekproef vormen uit een grotere populatie is het beoordelaars-effect "fixed". Omdat de werkstukken die beoordeeld worden een toevallige steekproef vormen uit een veel grotere populatie van werkstukken, is voor geval 3 een tweeweg "mixed" model passend. Dit model ziet er als volgt uit:

$$x_{ij} = \mu + a_i + b_j + (ab)_{ij} + e_{ij}$$

waarbij:
x_{ij} = het oordeel van beoordelaar i op het j-de werkstuk;
μ = het overall populatiegemiddelde van de beoordelingen;
a_i = het verschil van μ met de beoordelingen van de i-de beoordelaar;
b_j = het verschil van μ met de ware score (gemiddelde van herhaalde beoordelingen) op het j-de werkstuk;
$(ab)_{ij}$ = de mate waarin de i-de beoordelaar afwijkt van zijn gebruikelijke beoordelingstendens wanneer hij geconfronteerd wordt met het j-de werkstuk (interactie);
e_{ij} = de toevallige fout in de score van beoordelaar i op het j-de werkstuk.

De variantie-analyse voor het tweeweg mixed model wordt in tabel 5.3 gespecificeerd (Shrout en Fleiss, 1979).

Tabel 5.3: Variantie-analyse voor het tweeweg mixed model. (f = k(k-1); df = aantal vrijheidsgraden; MS = variantie; EMS = verwachte variantie)

bron	df	MS	EMS
werkstukken	n-1	MS_w	$k\sigma_w^2 + \sigma_e^2$
beoordelaars	k-1	MS_r	$n\sum r^2/(k-1) + f\sigma_I^2 + \sigma_e^2$
error	(n-1)(k-1)	MS_e	$f\sigma_I^2 + \sigma_e^2$
totaal	nk-1		

In geval 3 is het beoordelaars-effect fixed. Een implicatie hiervan is dat er geen zuivere schatter voor σ_w^2 (werkstukken-effect) beschikbaar is als sprake is van een interactie-effect ($\sigma_I^2 > 0$). De ICC voor geval 3 wordt gegeven door:

$$\rho = \frac{\sigma_w^2 - \sigma_I^2/(k-1)}{\sigma_w^2 + \sigma_I^2 + \sigma_e^2}$$

De consistente maar onzuivere schatter wordt gegeven door:

$$R = \frac{MS_w - MS_e}{MS_w + (k-1)MS_e}$$

Toetsing van de nulhypothese - dat $\rho = 0$ - gebeurt aan de hand van $F_0 = MS_w/MS_e$ met (n-1) en (n-1)(k-1) vrijheidsgraden.
Tot nu toe is alleen de ICC besproken die de betrouwbaarheid schat van één enkele beoordelaar. Soms gaat de belangstelling uit naar de betrouwbaarheid van de gemiddelde beoordeling van een aantal beoordelaars. Voor geval 3 geven Shrout en Fleiss (1979) een formule om een ICC te berekenen voor de gemiddelde beoordeling als er geen interactie-effect is tussen beoordelaars en werkstukken:

$$R = \frac{MS_w - MS_e}{MS_w}$$

De significantie-toets is identiek aan die welke gebruikt wordt bij het schatten van de betrouwbaarheid van één beoordelaar.

5.3.4 Index T

Tinsley en Weiss (1975) bespreken in hun overzicht van maten voor inter-beoordelaarsbetrouwbaarheid twee indices voor inter-beoordelaarsovereenstemming voor ordinale en intervalschalen, geformuleerd door respectievelijk Lawlis en Lu (1972) en Lu (1971). Volgens Tinsley en Weiss is Lu's index eerder een maat voor associatie dan voor overeenstemming; reden waarom deze index hier niet besproken zal worden. Lawlis en Lu suggereren de volgende nonparametrische chi-kwadraat als een test voor de significantie van inter-beoordelaarsovereenstemming:

$$\chi^2 = \frac{(N1 - NP - 0.5)^2}{NP} + \frac{(N2 - N(1 - P) - 0.5)^2}{N(1 - P)}$$

waarbij:
N1 = het aantal waarnemingen;
N = het aantal beoordeelde individuen;
P = de waarschijnlijkheid van kansovereenstemming met betrekking tot één individu;
0.5= continuïteitscorrectie;
N2 = het aantal niet-overeenstemmingen.

Deze index is verdeeld als chi-kwadraat met één vrijheidsgraad.

De test is alleen toepasbaar als de inter-beoordelaarsovereenstemming (N1) groter is dan de overeenstemming die verwacht kan worden op basis van kans alleen (NP). Het niet verkrijgen van een significante chi-kwadraat betekent dat de hypothese van "toevallig toegekende beoordelingen" niet verworpen kan worden. Een significante chi-kwadraat betekent dat de waargenomen overeenstemming groter is dan de overeenstemming die verwacht kan worden op basis van kans. Omdat de onderzoeker daarnaast ook graag informatie heeft over de mate van overeenstemming stellen Tinsley en Weiss de volgende index voor:

$$T = \frac{N1 - NP}{N - NP}$$

T is op dezelfde wijze gedefinieerd als coëfficiënt Kappa (zie paragraaf 5.3.2) en wordt alleen berekend als de hypothese van kansovereenstemming verworpen is. T is gelijk aan 0 als de waargenomen overeenstemming gelijk is aan de verwachte kansovereenstemming. T is gelijk aan 1 als de inter-beoordelaarsovereenstemming perfect is. Positieve T-waarden geven aan dat de waargenomen overeenstemming groter is dan de kansovereenstemming. Een belangrijk voordeel van de besproken chi kwadraat-maat is dat de onderzoeker zelf het criterium voor overeenstemming kan bepalen, waardoor het probleem van de "alles-of-niets" overeenstemming vermeden wordt. Zo is het mogelijk dat de onderzoeker overeenstemming definieert als beoordelingen die niet meer dan één schaalpunt van elkaar verschillen.
Bij het gebruik van Lawlis en Lu's chi-kwadraat dient men zich te realiseren dat deze maat veronderstelt dat elke schaalwaarde dezelfde waarschijnlijkheid heeft om gebruikt te worden, onder de hypothese dat de beoordelingen puur toevallig zijn. Door de algemene neiging van beoordelaars om extreme schaalwaarden te vermijden, zal aan deze assumptie vaak niet voldaan worden. De beperking van het aantal schaalpunten heeft tot gevolg dat de waarschijnlijkheid van kansovereenstemming onderschat wordt en dat de significantie van de waargenomen overeenstemming dus overschat wordt. Er zijn twee oplossingen voor dit probleem. De onderzoeker kan het vereiste significantieniveau aanpassen, door bijvoorbeeld te eisen dat de overschrijdingskansen kleiner of gelijk zijn aan 4% of 3% in plaats van de gebruikelijke 5%. Of de onderzoeker berekent de waarschijnlijkheid van kansovereenstemming op basis van minder schaalpunten dan in werkelijkheid aanwezig zijn.

5.4 Validiteit van metingen

In dit hoofdstuk is tot nu toe alleen gesproken over de betrouwbaarheid van beoordelingen. Maar betrouwbaarheid alleen is een onvoldoende waarborg voor kwalitatief goede beoordelingen. Ondanks het gebruik van een betrouwbare beoordelingsmethode kan het voorkomen dat twee verschillende beoordelingen geen verschil in

prestatie representeren. Hier wordt de validiteitsvraag actueel. De betrouwbaarheid is een noodzakelijke maar niet voldoende voorwaarde voor validiteit. De validiteit van de beoordeling gaat over de representativiteit van de beoordeling met betrekking tot het werkstuk. Ofwel: de accuraatheid (accuracy) van de beoordeling. Houpt en Kress (1973) definieerden het begrip "accuraatheid" operationeel als de mate van overeenstemming met expertoordelen. Zij noemden accuraatheid een vorm van inhoudsvaliditeit. In navolging van Houpt en Kress wordt in dit onderzoek de validiteit van de beoordelingen vastgesteld door de associatie en de overeenstemming te berekenen tussen beoordelingen en de referentie-oordelen. Voor deze berekeningen zal gebruik gemaakt worden van dezelfde associatie- en overeenstemmingsmaten als die besproken zijn in paragraaf 5.3.

VI RESULTATEN VAN HET ONDERZOEK NAAR HET FUNCTIONEREN VAN HET BEOORDELINGSPROTOCOL EN HET TRAININGSPROGRAMMA

6.1 Inleiding

Kan de kwaliteit van preklinische werkstukbeoordelingen verbeterd worden door beoordelaars gebruik te laten maken van een beoordelingsprotocol en door hen te trainen? In paragraaf 6.2 wordt deze centrale vraagstelling opgedeeld in drie meer concrete onderzoeksvragen, die vervolgens geoperationaliseerd worden door middel van deelvragen. In paragraaf 6.3 worden voor elke geformuleerde deelvraag de resultaten gepresenteerd en besproken. Met de conclusies in paragraaf 6.4 wordt het hoofdstuk afgesloten.

6.2 Vraagstellingen en operationalisaties

De nieuwe beoordelingsmethode onderscheidt zich van de in het preklinisch onderwijs gebruikte beoordelingsmethode door het gebruik van een beoordelingsprotocol. Een vanzelfsprekende vraag is natuurlijk of de nieuw ontwikkelde methode beter is dan de vigerende. Anders gezegd: kan meer vertrouwen gesteld worden in de beoordelingen aan de hand van het beoordelingsprotocol dan in de beoordelingen aan de hand van de kenmerkmethode? Dit leidt tot de formulering van de eerste vraagstelling:

1. Zijn beoordelingen aan de hand van het beoordelingsprotocol betrouwbaarder dan aan de hand van de kenmerkmethode?

Daarnaast is het belangrijk dat beoordelingen representatief zijn voor de beoordeelde werkstukken. Het is niet voldoende als een beoordelaar het eens is met zichzelf of met een andere beoordelaar over de kwaliteit van een bepaald werkstuk. De beoordeling moet tevens een nauwkeurige afspiegeling zijn van die kwaliteit. De tweede vraagstelling luidt daarom:

2. Zijn beoordelingen aan de hand van het beoordelingprotocol meer valide dan aan de hand van de kenmerkmethode?

Het beoordelingsprotocol en het ontwikkelde trainingsprogramma zijn geïntegreerd getest. De laatste onderzoeksvraag luidt derhalve:

3. Is er sprake van trainings-effecten?

Voordat tot beantwoording van de onderzoeksvragen kan worden overgegaan moet eerst elke vraagstelling geoperationaliseerd worden.

Operationalisatie van de eerste vraagstelling (betrouwbaarheid)

Inter-beoordelaarsovereenstemming is een schatter voor de betrouwbaarheid van beoordelingen (par. 5.2). De betrouwbaarheidsschattingen zijn nauwkeuriger naarmate ze op meer waarnemingen gebaseerd zijn. Daarom worden de berekeningen uitgevoerd tussen elke beoordelaar en alle overige beoordelaars, in plaats van tussen elk mogelijk beoordelaarspaar. De berekeningen worden uitgevoerd over de kenmerk- en subkenmerkbeoordelingen. De interesse gaat uit naar de betrouwbaarheid per werkstuk en per beoordelingsaspect. De betrouwbaarheidsberekening per werkstuk is interessant omdat beslissingen over studenten altijd genomen worden op basis van complete werkstukbeoordelingen en niet op basis van deelbeoordelingen van werkstukken. De betrouwbaarheidsberekening per beoordelingsaspect is van belang vanwege de informatie die dit oplevert over het functioneren van de afzonderlijke beoordelingsaspecten van beide beoordelingsmethoden.
De betrouwbaarheid kan ook geschat worden op basis van de overeenstemming die een beoordelaar weet te bereiken met zijn eigen, op een ander tijdstip gegeven, beoordelingen. Deze stabiliteitsberekeningen worden uitgevoerd over de zogenaamde "Herhaalwerkstukken" en het "Rode Draad" werkstuk (zie par. 4.5.5). Omdat niet elke trainee ieder werkstuk op alle subkenmerken beoordeeld heeft, is het niet mogelijk om de kenmerkmethode te vergelijken met de subkenmerkmethode met betrekking tot de betrouwbaarheid van cijfers, gebaseerd op respectievelijk kenmerk- en subkenmerkscores. Wel mogelijk is een vergelijking tussen cijfers die gebaseerd zijn op de kenmerkscores en zogenaamde "impressionistische" cijfers. Laatstgenoemde cijfers werden in de trainings-sessies door de beoordelaars aan de werkstukken toegekend, voorafgaand aan de beoordeling op kenmerken. In bijlage 5 wordt de betrouwbaarheid van beide cijfer-reeksen geschat door middel van de in paragraaf 5.3.3 en 5.3.4 besproken technieken.
De volgende deelvragen kunnen geformuleerd worden:

A. Hoe groot is de inter-beoordelaarsovereenstemming per werkstuk, op respectievelijk kenmerk- en subkenmerkniveau?
B. Hoe groot is de inter-beoordelaarsovereenstemming per beoordelingsaspect, op respectievelijk kenmerk- en subkenmerkniveau?
C. Hoe stabiel zijn de beoordelingen op respectievelijk kenmerk- en subkenmerkniveau?

Operationalisatie van de tweede vraagstelling (validiteit)

In paragraaf 5.4 werd beschreven dat de validiteit van beoordelingen geschat kan worden door de overeenstemming te berekenen die bereikt wordt met een kwaliteits-standaard (referentie-oordeel). Evenals bij de betrouwbaarheidsschattingen kan ook nu weer onderscheid gemaakt worden naar berekening per werkstuk en per beoordelingsaspect. Ook worden alle berekeningen weer apart uitgevoerd over de kenmerk- en de subkenmerkbeoordelingen. De deelvragen met betrekking tot de tweede vraagstelling luiden:

D. Hoe groot is de overeenstemming met het referentie-oordeel op respectievelijk kenmerk- en subkenmerkniveau, berekend per werkstuk?
E. Hoe groot is de overeenstemming met het referentie-oordeel op respectievelijk kenmerk- en subkenmerkniveau, berekend per beoordelingsaspect?

Operationalisatie van de derde vraagstelling (trainings-effect)

Per trainings-sessie wordt voor elke beoordelaar berekend hoe groot de overeenstemming is met het referentie-oordeel en met de overige beoordelaars. De berekeningen worden uitgevoerd over alle werkstukken die in de betreffende trainings-sessie zijn aangeboden. Om na te kunnen gaan of de training een verschillend effect heeft op de twee gebruikte beoordelingsmethoden, worden de overeenstemmingen bepaald voor zowel de kenmerk- als de subkenmerkmethode.
Een andere belangrijke variabele is de hoeveelheid tijd die training van beoordelaars in beslag neemt. Metingen tijdens de trainings-sessies lieten zien, dat beoordelaars gemiddeld twee minuten tijd nodig hadden voor het beoordelen van een werkstuk op kenmerkniveau. Dit gold zowel voor de eerste als voor de laatste trainings-sessie. Training beïnvloedde de benodigde beoordelingstijd op kenmerkniveau dus niet. Doordat steeds de tijdstippen werden genoteerd waarop beoordelaars met de training aanvingen en eindigden en door kennis van de gemiddelde beoordelingstijd op kenmerkniveau, kan voor elke beoordelaar bepaald worden hoeveel tijd gemiddeld besteed werd aan één subkenmerk-beoordeling. Het uitvoeren van deze berekeningen per trainings-sessie levert informatie op over de invloed van training op de benodigde beoordelingstijd per subkenmerk.
De te beantwoorden deelvragen luiden als volgt:
F. Is er sprake van een systematisch toenemende beoordelaarsovereenstemming naarmate aan meer trainings-sessies is deelgenomen?
G. Neemt de benodigde beoordelingstijd voor een subkenmerk systematisch af als gevolg van training?

6.3 Resultaten

6.3.1 Inleiding

Hoe sterk conclusies uit empirisch onderzoek zijn hangt voor een groot deel af van de opzet van het onderzoek. Een design, dat niet-relevante variabelen kan controleren door hun invloed te elimineren of constant te houden, biedt betere mogelijkheden voor "harde" conclusies dan een design, waarin deze controle niet mogelijk is. Experimentele en quasi-experimentele designs zijn echter in het onderwijs vaak niet realiseerbaar. Gebrek aan tijd en proefpersonen zijn de belangrijkste oorzaken. Zo ook in het onderhavige onderzoek, waarin het aantal proefpersonen beperkt bleef tot vijf. Om toch informatie te krijgen over de relatieve

kwaliteit van beide beoordelingsmethoden, alsmede over de invloed van beoordelaarstraining daarop, werd voor de in paragraaf 4.5 beschreven opzet gekozen.
In paragraaf 4.5.3 werden drie argumenten genoemd voor het gelijktijdig gebruik van de kenmerk- en subkenmerkmethode in de trainings-sessies. Helaas zijn er ook nadelen aan deze constructie verbonden. Zo kan bijvoorbeeld niet beschikt worden over volledige waarnemingen op subkenmerkniveau; een gevolg van het feit dat beoordeling op dit niveau afhankelijk is gesteld van de beoordeling op kenmerkniveau. Alleen als op kenmerkniveau geen overeenstemming wordt bereikt met het referentie-oordeel moet op subkenmerkniveau beoordeeld worden.
Een ander nadeel van het voorwaardelijk gebruik van de subkenmerkmethode is, dat het niet goed mogelijk is om vast te stellen of eventuele positieve resultaten (hogere overeenstemmingen op subkenmerkniveau) het gevolg zijn van de structurerende werking van de beoordelingsprotocollen en/of van de meer kritische instelling van de beoordelaars, als resultaat van de terugkoppeling op kenmerkniveau. In een poging om een meer directe vergelijking te kunnen maken tussen de kenmerk- en de subkenmerkmethode, werd een maand na afloop van de trainings-sessies opnieuw een beroep gedaan op de trainees. Hun werd gevraagd om zes toevallig uit het werkstukkenbestand getrokken werkstukken op alle subkenmerken te beoordelen. Deze werkstukken werden dus op subkenmerkniveau beoordeeld, zonder voorafgaande kennis van de kwaliteit op kenmerkniveau. De op deze beoordelingen gebaseerde overeenstemmingen kunnen direct vergeleken worden met de uit de trainings-sessies stammende overeenstemmingen op kenmerkniveau (zie bijlage 4). De gegevens uit deze extra sessie vormen een aanvulling op de gegevens uit de originele trainings-sessies (zie bijlage 6) en kunnen de interpretatie vereenvoudigen van eventuele verschillen in beoordelaarsovereenstemming tussen kenmerk- en subkenmerkmethode. Op daarvoor geschikte momenten zullen ze gepresenteerd worden onder het kopje "directe vergelijking kenmerk- en subkenmerkmethode". Door langdurige afwezigheid kon één trainee niet deelnemen aan deze extra trainings-sessie. In de tabellen waarin de beoordelingsresultaten van de extra trainings-sessie worden gepresenteerd zal beoordelaar 3 daarom ontbreken.

Op deze plaats moet nog een opmerking gemaakt worden over de toetsing van coëfficiënt Kappa. Doordat alleen op subkenmerkniveau beoordeeld werd als op kenmerkniveau geen overeenstemming bereikt was met het referentie-oordeel, varieert het aantal beoordelingen op subkenmerkniveau zeer sterk van werkstuk tot werkstuk en van beoordelaar tot beoordelaar. Als gevolg daarvan kunnen de Kappa coëfficiënten die over de subkenmerkbeoordelingen berekend zijn niet statistisch getoetst worden. De Kappa's berekend over de beoordelingsscores uit de extra trainingssessie, daarentegen, kunnen wel getoetst worden. In die sessies werd ieder werkstuk door elke beoordelaar op alle subkenmerken beoordeeld, zodat het aantal beoordelingen per werkstuk en per beoordelingsaspect steeds constant bleef. Op basis van dit constante aantal waarnemingen is toetsing van Kappa mogelijk.

6.3.2 Betrouwbaarheid van de kenmerk- en de subkenmerkmethode

6.3.2.1 Beantwoording van deelvraag A: Hoe groot is de inter-beoordelaarsovereenstemming per werkstuk op respectievelijk kenmerk- en subkenmerkniveau?

De vraag of werkstukken op subkenmerkniveau betrouwbaarder beoordeeld worden dan op kenmerkniveau, kan beantwoord worden door frequentieverdelingen op te stellen voor beide beoordelingsmethoden met betrekking tot de bereikte inter-beoordelaarsovereenstemmingen. Per overeenstemmingsklasse wordt het aantal werkstukken geturfd. De verwachting is, dat op subkenmerkniveau meer werkstukken in de hogere overeenstemmingsklassen zullen vallen dan op kenmerkniveau. Tabel 6.1 bevestigt dit vermoeden.

Tabel 6.1: Frequentieverdeling van de bereikte inter-beoordelaarsovereenstemming (Kappa) per beoordelaar (1 t/m 5) op kenmerk- (K) en subkenmerkniveau (SK). De waarden in de cellen representeren het aantal werkstukken.

	beoordelaars									
	1		2		3		4		5	
Kappa	K	SK	K	SK	K	SK	K	SK	K	SK
-.50 - -.40		1				2				
-.39 - -.30					1					
-.29 - -.20								1		
-.19 - -.10					1		1		1	
-.09 - .00	1		1		2		2	1	1	
.01 - .10	5				3		2	1	3	
.11 - .20	5	1	11	3	3		7		6	
.21 - .30	8	2	7	1	9	1	10	4	7	3
.31 - .40	14	6	8	3	14	6	16	4	6	4
.41 - .50	8	7	15	5	7	7	4	9	14	9
.51 - .60	1	3	1	9	5	7	3	7	6	7
.61 - .70	5	12	4	14	2	12	2	15	3	12
.71 - .80		14		9		9		5		11
.81 - .90				3		2				
.91 - 1.00						1				
totaal	47	46	47	47	47	47	47	47	47	46

Bij elke beoordelaar kan geconstateerd worden dat de frequentieverdeling voor de overeenstemming op subkenmerkniveau, ten opzichte van de overeenstemming op kenmerkniveau, is opgeschoven in

de richting van de hogere overeenstemmingsklassen. Tenzij anders vermeld, is de inter-beoordelaarsovereenstemming steeds berekend tussen de beoordelingen van een bepaalde beoordelaar en de beoordelingen van alle overige beoordelaars. Het verschil tussen beide beoordelingsmethoden tekent zich duidelijker af als gelet wordt op het percentage werkstukken dat boven een bepaalde grens (bijvoorbeeld Kappa > 0.60) ligt. In tabel 6.2 wordt voor de kenmerk- en subkenmerkmethode het percentage werkstukken gegeven met een Kappa coëfficiënt groter dan 0.60.

Tabel 6.2: Percentage werkstukken met een Kappa groter dan 0.60 op kenmerk-(K) en subkenmerkniveau (SK), uitgesplitst naar beoordelaars.

beoordelaars	beoordelingsmethode kenmerken	subkenmerken
1	11	57
2	9	55
3	4	51
4	4	43
5	6	50

De verschillen tussen de beoordelingsmethoden zijn groot en, belangrijker, doen zich bij elke beoordelaar in ongeveer dezelfde mate voor.

Een andere manier om verschillen tussen de kenmerk- en subkenmerkmethode te beschrijven, is berekening van de gemiddelde overeenstemming over alle werkstukken, op respectievelijk kenmerk- en subkenmerkniveau (zie tabel 6.3). De Kappa-waarden in tabel 6.3 (en in alle volgende tabellen) zijn gemiddelde Kappa's. De ene keer wordt de gemiddelde Kappa berekend per beoordelaar, de andere keer per beoordelingsaspect, per werkstuk of per sessie. Berekening van de gemiddelde Kappa geschiedt aan de hand van de volgende formule:

$$\bar{K}_L = \frac{\bar{P}_o - 1/L}{1 - 1/L}$$

waarbij: $\bar{P}_o = \sum f_o / K*N$ (proportie overeenstemming);
$\sum f_o$ = aantal waargenomen overeenstemmingen;

N = aantal beoordelingen per beoordelaar;
K = aantal beoordelaars;
L = lengte van de beoordelingsschaal.

Ter illustratie: de waarde 0.34 in tabel 6.3 is de gemiddelde, voor kans gecorrigeerde, overeenstemming tussen beoordelaar 1 en alle andere beoordelaars op kenmerkniveau. Met behulp van bovenstaande formule is voor elk beoordeeld werkstuk de overeenstemming berekend. Daarna zijn alle $\overline{\kappa}_L$'s opgeteld en is de som gedeeld door het aantal beoordeelde werkstukken.

Tabel 6.3: Gemiddelde ($\overline{X}$) inter-beoordelaarsovereenstemming per beoordelaar (Kappa) en standaardafwijking (s.d.), berekend over alle werkstukken op kenmerk- (K) en subkenmerkniveau (SK).

	beoordelingsmethode			
	kenmerken		subkenmerken	
beoordelaars	$\overline{X}$	s.d.	$\overline{X}$	s.d.
1	.34	.18	.54	.25
2	.36	.16	.59	.18
3	.31	.22	.53	.29
4	.30	.17	.50	.22
5	.36	.19	.56	.16

Uit tabel 6.3 kan geconcludeerd worden dat elke beoordelaar een aanzienlijk hogere overeenstemming bereikt met alle overige beoordelaars als werkstukken op subkenmerken beoordeeld worden in plaats van op kenmerken.

Directe vergelijking kenmerk- en subkenmerkmethode

In par. 6.3.1 is uiteengezet dat, waar mogelijk, een directe vergelijking gemaakt wordt tussen de beoordelingen op kenmerk- en subkenmerkniveau. Dit houdt in, dat voor beide beoordelingsmethoden over volledige waarnemingen beschikt wordt en dat de beoordelingen op subkenmerkniveau niet beïnvloed kunnen zijn door voorafgaande beoordelingen op kenmerkniveau. Tabel 6.4 bevat de reële overeenstemmingen (Kappa) per werkstuk, berekend tussen elke beoordelaar en alle overige beoordelaars op respectievelijk kenmerk- en subkenmerkniveau.

Tabel 6.4: Inter-beoordelaarsovereenstemming (Kappa) op kenmerk- (K) en subkenmerkniveau (SK), berekend per werkstuk en per beoordelaar (1, 2, 4 en 5). (directe vergelijking)

	werkstukken											
	36		128		374		449		658		868	
	K	SK	K	SK	K	SK	K	SK	K	SK	K	SK
1	-.17	.74**	.08	.67**	.67	.68*	.58*	.68**	.33	.77**	.08	.67**
2	.00	.74**	.00	.67**	.50	.65**	.17	.71**	.50	.58*	.17	.71**
4	.00	.69**	.25	.53*	.67*	.68**	.33	.71**	.33	.70**	.25	.71**
5	.00	.76**	.00	.78**	.50	.71**	.58*	.74**	.50	.67**	-.17	.78**

* = $p<.05$
** = $p<.01$

De Kappa's op subkenmerkniveau zijn allemaal statistisch significant. Dit betekent dat de kans maximaal vijf procent is, dat de gevonden waarden toevallig groter zijn dan nul. Op kenmerkniveau moet voor vier van de zes werkstukken geconstateerd worden dat geen enkele beoordelaar een reële overeenstemming heeft bereikt die statistisch significant is.

Tabel 6.5: Gemiddelde ($\bar{X}$) inter-beoordelaarsovereenstemming per beoordelaar (Kappa) en standaardafwijking (s.d.) op kenmerk- (K) en subkenmerkniveau (SK). (directe vergelijking)

	beoordelingsmethode			
	kenmerken		subkenmerken	
beoordelaars	$\bar{X}$	s.d.	$\bar{X}$	s.d.
1	.26	.32	.70	.04
2	.22	.23	.68	.06
4	.31	.22	.67	.07
5	.24	.33	.74	.04

Tabel 6.4 kan overzichtelijker weergegeven worden door over de zes werkstukken de gemiddelde Kappa per beoordelaar te berekenen (zie tabel 6.5). De verschillen tussen de kenmerk- en de subkenmerkmethode met betrekking tot de inter-beoordelaarsovereenstemming (Kappa) zijn in tabel 6.5 (directe vergelijking) aanmerkelijk groter dan in tabel 6.3 (resultaten originele trainings-sessies). Op grond van deze resultaten kan geconcludeerd worden, dat de in tabel 6.3 gesignaleerde verschillen tussen kenmerk- en subkenmerkmethode het gevolg zijn van verschillen in betrouwbaarheid van genoemde beoordelingsmethoden.

6.3.2.2 Beantwoording van deelvraag B: Hoe groot is de inter-beoordelaarsovereenstemming per beoordelingsaspect, op respectievelijk kenmerk- en subkenmerkniveau?

Worden de diverse aspecten waarop werkstukken gekwalificeerd worden, betrouwbaarder beoordeeld met de subkenmerk- dan met de kenmerkmethode? De kwaliteit van een beoordeling op een bepaald kenmerk wordt vergeleken met de kwaliteit van de beoordelingen op de subkenmerken, die operationalisaties van het betreffende kenmerk zijn. De in par. 6.3.1 genoemde beperkte mogelijkheid tot vergelijking van beide beoordelingsmethoden, is met name van toepassing als gekeken wordt naar de verschillen per beoordelingsaspect. De subkenmerkbeoordelingen die vergeleken worden met de kenmerkbeoordelingen zijn nu uitsluitend afkomstig van beoordelaars die op kenmerkniveau geen overeenstemming wisten te bereiken met het referentie-oordeel. Dit gegeven kan de overeenstemming op subkenmerkniveau systematisch beïnvloed hebben. De resultaten van de directe vergelijking tussen de kenmerk- en de subkenmerkmethode zijn daarom van extra groot belang.
Tabel 6.6 geeft de inter-beoordelaarsovereenstemmingen weer (uitgedrukt in coëfficiënt Kappa) die berekend zijn per beoordelingsaspect en per beoordelaar op respectievelijk kenmerk- en subkenmerkniveau. De tabel laat zien dat de overeenstemming tussen beoordelaars op subkenmerkniveau meestal groter is dan op kenmerkniveau. Uitzonderingen hierop vormen de beoordelingsaspecten "afwerking" (AF) en (in mindere mate) "pulpo axiale afschuining" (PA). Kennelijk laat de voorgeschreven beoordelingsmethode voor "afwerking" ook op subkenmerkniveau te veel ruimte voor interpretatie. Voor het beoordelingsaspect "pulpo axiale afschuining" is slechts één subkenmerk beschikbaar. De beoordelaar moet volgens de criterium-omschrijving van dit subkenmerk in staat zijn om onderscheid te maken tussen 0.3 en 0.8 mm. Deze afmetingen zijn dermate klein dat redelijkerwijs geen hoge overeenstemming tussen beoordelaars verwacht mag worden. "Diepte" (DI) blijkt een beoordelingsaspect dat op kenmerkniveau erg lastig te beoordelen is. Waarschijnlijk is dit een gevolg van het feit dat dit aspect, meer dan andere aspecten, op een groot aantal plaatsen in de preparatie gemeten kan worden. Op subkenmerkniveau is de overeenstemming met betrekking tot diepte veel groter, hoewel toch wat achterblijvend ten opzichte van aspecten als "outline" (OU), "caviteit-oppervlakte hoek" (CA) en "convergentie" (CO).

Tabel 6.6: Inter-beoordelaarsovereenstemming (Kappa) per beoordelingsaspect en per beoordelaar op kenmerk- (K) en subkenmerkniveau (SK). Berekeningen over alle werkstukken. Onder de stippellijn: gemiddelde ($\bar{X}$) inter-beoordelaarsovereenstemming per beoordelingsaspect en bijbehorende standaardafwijking (s.d.).

	beoordelingsaspecten											
	OU		DI		CA		CO		PA		AF	
beoord.	K	SK	K	SK	K	SK	K	SK	K	SK	K	SK
1	.40	.71	.23	.58	.35	.72	.35	.67	.30	.54	.42	.42
2	.43	.67	.18	.58	.19	.71	.42	.62	.35	.50	.33	.31
3	.48	.73	.14	.52	.29	.70	.42	.65	.24	.35	.30	.41
4	.33	.66	.11	.55	.38	.58	.34	.61	.25	.55	.33	.41
5	.47	.70	.20	.57	.34	.69	.39	.66	.34	.42	.40	.43
$\bar{X}$	.42	.69	.17	.56	.31	.68	.38	.64	.30	.47	.36	.40
s.d.	.06	.03	.05	.03	.07	.06	.04	.03	.05	.09	.05	.05

Directe vergelijking kenmerk- en subkenmerkmethode

De resultaten van de directe vergelijking tussen de twee beoordelingsmethoden in tabel 6.7, zijn een indicatie voor de mate waarin staat gemaakt kan worden op de resultaten van tabel 6.6. Op één na zijn alle Kappa coëfficiënten op subkenmerkniveau statistisch significant op minimaal het vijf procent toetsingsniveau. Op kenmerkniveau zijn slechts twee Kappa's statistisch significant. De betekenis hiervan is, dat op kenmerkniveau niet gesproken kan worden van echte overeenstemming tussen beoordelaars, vanwege de grote kans op toeval, terwijl op subkenmerkniveau die kans bijna nergens groter is dan vijf procent.

Tabel 6.7: Inter-beoordelaarsovereenstemming (Kappa) op kenmerk- (K) en subkenmerkniveau (SK) per beoordelingsaspect en per beoordelaar (1, 2, 4 en 5). (directe vergelijking)

	beoordelingsaspecten											
	OU		DI		CA		CO		PA		AF	
	K	SK	K	SK	K	SK	K	SK	K	SK	K	SK
1	.33	.74**	.17	.58**	.25	.83**	.42	.79**	.00	.58**	.42	.61**
2	.50	.74**	.00	.72**	.25	.77**	.17	.73**	-.17	.75**	.58	.46*
4	.50	.71**	.08	.72**	.42	.70**	.42	.64**	.00	.75**	.42	.60**
5	.67*	.70**	-.25	.75**	.25	.77**	.33	.76**	.00	.42	.42	.69**

* = $p<.05$
** = $p<.01$

In tabel 6.8 worden de resultaten van tabel 6.7 samengevat door middel van de gemiddelde overeenstemming (Kappa) per beoordelingsaspect. De eveneens opgenomen standaardafwijkingen zijn een indicatie voor de mate waarin de Kappa's van beoordelaar tot beoordelaar verschillen. De gemiddelde overeenstemming tussen beoordelaars is voor alle beoordelingsaspecten groter als werkstukken op subkenmerken beoordeeld worden. De verschillen in overeenstemming tussen de kenmerk- en de subkenmerkmethode zijn over het algemeen groot, het beoordelingsaspect "afwerking" uitgezonderd. Op kenmerkniveau vallen de zeer lage overeenstemmingen op die beoordelaars bereiken op de aspecten "diepte" en "pulpo axiale afschuining". In de resultaten van de originele trainingssessies (zie tabel 6.6) kan hetzelfde verschijnsel geconstateerd worden. De conclusie luidt, dat de beoordelingsresultaten uit de extra sessie de, in tabel 6.6, geconstateerde verschillen tussen kenmerk- en subkenmerkmethode bevestigen en dat die verschillen voornamelijk veroorzaakt worden door de gehanteerde beoordelingsmethode en niet (of in mindere mate) door het voorwaardelijk gebruik van de subkenmerkmethode.

Tabel 6.8: Gemiddelde ($\bar{X}$) inter-beoordelaarsovereenstemming per beoordelingsaspect (Kappa) en standaardafwijkingen (s.d.) op kenmerk-(K) en subkenmerkniveau (SK). (directe vergelijking)

	beoordelingsmethode			
	kenmerken		subkenmerken	
beoordelingsaspecten	$\bar{X}$	s.d.	$\bar{X}$	s.d.
outline	.50	.14	.75	.03
diepte	.00	.18	.69	.08
cav.opp.hoek	.29	.09	.77	.05
convergentie	.34	.12	.73	.06
pul.ax.afsch.	-.04	.09	.63	.16
afwerking	.46	.08	.59	.10

6.3.2.3 Beantwoording van deelvraag C: Hoe stabiel zijn beoordelingen op respectievelijk kenmerk- en subkenmerkniveau?

De intra-beoordelaarsbetrouwbaarheid wordt geschat door middel van overeenstemmingsberekeningen tussen de beoordelingen van een beoordelaar op een bepaald moment met zijn beoordelingen op een ander moment. Deze berekeningen worden uitgevoerd over de zogenoemde "Herhaalwerkstukken" en het "Rode Draad" werkstuk. Een Herhaalwerkstuk is een willekeurig uit het werkstukkenbestand getrokken werkstuk, dat twee keer (met een tussentijd van ongeveer twee weken) aan een beoordelaar is aangeboden. Het Rode Draad werkstuk is zes keer door elke beoordelaar beoordeeld, eveneens met een tussentijd van telkens ongeveer twee weken. De overeenstemming met betrekking tot het Rode Draad werkstuk wordt berekend tussen de eerste beoordeling en de vijf daarop volgende beoordelingen van dit werkstuk. In tabel 6.9 wordt een overzicht gegeven van de stabiliteit met betrekking tot de Herhaalwerkstukken en het Rode Draad werkstuk. Uit deze tabel blijkt dat alle beoordelaars stabieler beoordelen aan de hand van de subkenmerkmethode dan aan de hand van de kenmerkmethode. Dit geldt voor zowel de Herhaalwerkstukken als voor het Rode Draad werkstuk.

Tabel 6.9: Intra-beoordelaarsovereenstemmingen (Kappa) per beoordelaar, berekend over de Herhaalwerkstukken en het Rode Draad werkstuk op kenmerk- (K) en subkenmerkniveau (SK). Onder de stippellijn: gemiddelde ($\bar{X}$) intra-beoordelaarsovereenstemmingen en standaardafwijkingen (s.d.).

beoordelaars	Herhaalwerkstukken K	Herhaalwerkstukken SK	Rode Draad werkstuk K	Rode Draad werkstuk SK
1	.36	.69	.35	.93
2	.36	.55	.25	.47
3	.32	.61	.15	.65
4	.46	.68	.40	.68
5	.64	.69	.10	.61
$\bar{X}$	.43	.64	.25	.67
s.d.	.13	.06	.13	.17

De over alle beoordelaars berekende gemiddelde intra-beoordelaarsovereenstemming is op subkenmerkniveau ongeveer gelijk voor de Herhaalwerkstukken en het Rode Draad werkstuk. Op kenmerkniveau, echter, is de intra-beoordelaarsovereenstemming voor het Rode Draad werkstuk aanzienlijk kleiner dan voor de Herhaalwerkstukken. Vermoedelijk is dit een gevolg van het feit dat het Rode Draad werkstuk zes keer door elke beoordelaar beoordeeld is, waardoor de kans op niet-overeenstemming natuurlijk groter wordt dan bij een eenmalige herhaalbeoordeling. Dat een dergelijk verschil zich niet voordoet op subkenmerkniveau, kan worden opgevat als een aanwijzing voor de grotere objectiviteit van de subkenmerkmethode. Opvallend is verder het verschil in spreiding op subkenmerkniveau tussen de overeenstemmingen berekend over de Herhaalwerkstukken en het Rode Draad werkstuk. De standaardafwijking rond de gemiddelde intra-beoordelaarsovereenstemming van het Rode Draad werkstuk is bijna drie keer zo groot als die van de Herhaalwerkstukken. De vrij extreme overeenstemmingen die beoordelaar 1 (0.93) en beoordelaar 2 (0.47) met zichzelf bereiken zijn de oorzaak hiervan.

6.3.2.4 Discussie

Over het algemeen geven de resultaten aanleiding om te concluderen dat aangeboden werkstukken betrouwbaarder beoordeeld worden met de subkenmerk- dan met de kenmerkmethode. De Kappa coëfficiënten die berekend werden over de subkenmerkbeoordelingen zijn gemiddeld 63 procent groter dan de Kappa's over de kenmerkbeoordelingen (zie tabel 6.3). Voor de werkstukken die direct met elkaar vergeleken kunnen worden is dit verschil veel groter, gemiddeld 176 procent (zie tabel 6.5). Opgemerkt moet hier worden, dat dit laatste cijfer gebaseerd is op een steekproef van slechts zes werkstukken, waardoor de kans op toevallige verschillen vrij groot is. De tamelijk grote standaardafwijkingen rond de gemiddelde Kappa coëfficiënten voor de beoordelingen op subkenmerkniveau (zie tabel 6.3) geven aan dat, ondanks de gedetailleerde criteria, toch nog sprake is van grote verschillen in de overeenstemming. Kennelijk is overeenstemming niet even gemakkelijk te bereiken over elk werkstuk. Hier dient zich het onderscheid aan tussen eenvoudig en moeilijk te beoordelen werkstukken. Vermoedelijk kunnen werkstukken die duidelijk goed of slecht zijn als "eenvoudig" gekenschetst worden. Moeilijk te beoordelen werkstukken zijn de zogenaamde "grensgevallen". Hieraan dient veel aandacht besteed te worden, aangezien een groot deel van de in het preklinisch onderwijs vervaardigde klasse II-tweevlakspreparaties met een 5 of een 6 (tienpuntsschaal) gekwalificeerd wordt. In studiejaar 1982-1983 werd 47 procent van alle klasse II-tweevlakspreparaties met een 5 of een 6 gewaardeerd. Bij "grensgevallen" kunnen verkeerde beoordelingen snel leiden tot ten onrechte genomen zak-/slaagbeslissingen. Het verdient daarom aanbeveling om in toekomstige trainingen extra aandacht te besteden aan het beoordelen van "grensgevallen".

Als naar de beoordelingsaspecten afzonderlijk wordt gekeken, kan geconstateerd worden dat de overeenstemming tussen beoordelaars meestal veel groter is bij gebruik van de subkenmerkmethode (zie tabel 6.6, 6.7 en 6.8). Een uitzondering hierop vormen de aspecten "pulpo axiale afschuining" en "afwerking". Voor het eerstgenoemde aspect lijkt de oorzaak te liggen in het feit dat een met het blote oog nauwelijks zichtbare eigenschap van een preparatie beoordeeld moet worden en dat voor dit aspect slechts één subkenmerk geformuleerd is. Voor het aspect "afwerking" moet vastgesteld worden dat de omschrijvingen van de criteria niet voldoende objectief zijn geweest om dit moeilijke aspect betrouwbaar te kunnen beoordelen.

Voor het preklinisch onderwijs impliceert beoordelen op subkenmerken een verbeterde terugkoppeling naar de student over zijn prestaties. In de eerste plaats omdat de betrouwbaarheid van de beoordelingen groter is dan bij beoordeling op kenmerken. Een student kan met meer vertrouwen de voortgang van het leerproces baseren op de informatie die de beoordeling van zijn prestaties oplevert.

In de tweede plaats omdat de subkenmerkscores de student meer gedetailleerde informatie bieden over de aard van de gesignaleerde tekortkomingen. Op grond daarvan kan de student gericht werken aan verbetering van zijn vaardigheid. Daarnaast gaat een preventieve werking uit van het beoordelingsprotocol omdat het een leidraad kan zijn bij het vervaardigen van preparaties. Kennis van de beoordelingscriteria stuurt het leerproces. Om die reden verdient het aanbeveling het beoordelingsprotocol aan de studenten ter beschikking te stellen als onderdeel van het cursusmateriaal.

6.3.3 Validiteit van werkstukbeoordelingen

6.3.3.1 Beantwoording van deelvraag D: Hoe groot is de overeenstemming met het referentie-oordeel op respectievelijk kenmerk- en subkenmerkniveau, berekend per werkstuk?

De validiteit van beide beoordelingsmethoden wordt geschat via de overeenstemming tussen de scores van de beoordelaars en het referentie-oordeel. Tabel 6.10 geeft de gemiddelde overeenstemming per werkstuk die beoordelaars bereikt hebben met het referentie-oordeel. De overeenstemming met het referentie-oordeel is op subkenmerkniveau bij bijna alle beoordelaars twee keer zo groot als op kenmerkniveau. De standaardafwijkingen rond de Kappa's zijn vrij groot voor beide beoordelingsmethoden. Maar relatief gezien zijn de standaardafwijkingen op subkenmerkniveau kleiner, hetgeen betekent dat bij beoordeling van werkstukken met de subkenmerkmethode, minder gemakkelijk onderscheid gemaakt kan worden in moeilijk en makkelijk te beoordelen werkstukken.

Directe vergelijking kenmerk- en subkenmerkmethode

Tabel 6.11 bevat overeenstemmingen met het referentie-oordeel op kenmerk- en subkenmerkniveau die direct met elkaar vergeleken kunnen worden, omdat ze berekend zijn op basis van volledige gegevens en niet vertekend kunnen zijn door de opzet van de trainings-sessies. In deze tabel is slechts één Kappa coëfficiënt significant op kenmerkniveau. Uiteraard is dit mede een gevolg van het kleine aantal waarnemingen (n=6) op kenmerkniveau. Op subkenmerkniveau, daarentegen, zijn bijna alle Kappa's statistisch significant op minstens het vijf procent toetsingsniveau.

Tabel 6.10: Gemiddelde ($\bar{X}$) overeenstemming (Kappa) met het referentie-oordeel en standaardafwijking (s.d.), berekend per beoordelaar over alle werkstukken op kenmerk- (K) en subkenmerkniveau (SK).

	beoordelingsmethode			
	kenmerken		subkenmerken	
beoordelaars	$\bar{X}$	s.d.	$\bar{X}$	s.d.
1	.26	.30	.55	.26
2	.23	.29	.55	.30
3	.28	.29	.49	.35
4	.21	.34	.44	.25
5	.28	.29	.50	.23

Tabel 6.11: Overeenstemmingen (Kappa) met het referentie-oordeel, berekend per beoordelaar (1, 2, 4 en 5) en per werkstuk op kenmerk- (K) en subkenmerkniveau (SK). (directe vergelijking)

	werkstukken											
	36		128		374		449		658		868	
	K	SK	K	SK	K	SK	K	SK	K	SK	K	SK
1	.00	.53*	.25	.95**	.50	.71**	.25	.61*	.50	.77**	.25	.48
2	.00	.69*	.50	.74**	.25	.56*	.50	.71**	.50	.58*	.50	.43
4	-.25	.43	.50	.53*	.50	.66**	-.25	.71**	.50	.58*	.75*	.69**
5	-.25	.53*	.25	.90**	.50	.61*	.25	.61**	.25	.63*	.25	.53*

* = $p<.05$
** = $p<.01$

Tabel 6.12 vat tabel 6.11 samen door de gemiddelde overeenstemming te geven die elke beoordelaar bereikt heeft met het referentie-oordeel.

Tabel 6.12: Gemiddelde ($\bar{X}$) overeenstemming (Kappa) met het referentie-oordeel en standaardafwijking (s.d.) op kenmerk-(K) en subkenmerkniveau(SK), uitgesplitst naar beoordelaars. (directe vergelijking)

	beoordelingsmethode			
	kenmerken		subkenmerken	
beoordelaars	$\bar{X}$	s.d.	$\bar{X}$	s.d.
1	.29	.19	.68	.17
2	.38	.21	.62	.12
4	.29	.43	.60	.11
5	.21	.25	.64	.14

De Kappa's vertonen grote en systematische verschillen tussen de twee beoordelingsmethoden; de overeenstemming is veel groter als de werkstukken aan de hand van de subkenmerkmethode beoordeeld worden. Deze resultaten komen overeen met de resultaten uit de originele trainings-sessies (tabel 6.10) en rechtvaardigen de conclusie, dat klasse II-tweevlakspreparaties meer valide beoordeeld worden met de subkenmerk- dan met de kenmerkmethode.

6.3.3.2 Beantwoording van deelvraag E: Hoe groot is de overeenstemming met het referentie-oordeel op respectievelijk kenmerk- en subkenmerkniveau, berekend per beoordelingsaspect?

Geven subkenmerk-beoordelingen beter de kwaliteit weer van een werkstuk op een bepaald aspect dan kenmerk-beoordelingen? Om deze vraag te beantwoorden wordt per beoordelingsaspect een vergelijking gemaakt tussen de kenmerk- en subkenmerkmethode met betrekking tot de overeenstemming met het referentie-oordeel. In tabel 6.13 worden deze vergelijkingen voor elke beoordelaar weergegeven. Voor iedere beoordelaar geldt, dat de verschillen in overeenstemming met het referentie-oordeel tussen kenmerk- en subkenmerkmethode groot zijn en, behalve bij het beoordelingsaspect "afwerking", steeds in het voordeel van de subkenmerkmethode. Met betrekking tot het beoordelen van de afwerking moet

vermeld worden dat beoordelaars onderling soms perfecte overeenstemming bereikten, terwijl ze het tegelijkertijd oneens waren met het referentie-oordeel. Tabel 6.6 demonstreert echter, dat zeer voorzichtig moet worden omgegaan met uitspraken over de kwaliteit van de referentie-oordelen. Onderling blijken dezelfde beoordelaars ook veel moeite te hebben om overeenstemming te bereiken over de afwerking op subkenmerkniveau.

Tabel 6.13: Overeenstemming (Kappa) met het referentie-oordeel, berekend over alle werkstukken per beoordelingsaspect en per beoordelaar op respectievelijk kenmerk- (K) en subkenmerkniveau (SK). Onder de stippellijn: gemiddelde ($\bar{X}$) overeenstemming met het referentie-oordeel en standaardafwijking (s.d.).

	beoordelingsaspecten											
	OU		DI		CA		CO		PA		AF	
beoordelaars	K	SK	K	SK	K	SK	K	SK	K	SK	K	SK
1	.30	.62	.33	.61	.23	.79	.17	.67	.27	.46	.33	.37
2	.27	.68	.11	.59	.20	.74	.17	.68	.20	.42	.43	.17
3	.43	.58	.30	.58	.30	.80	.14	.61	.20	.63	.27	.23
4	.20	.59	.01	.55	.39	.57	.17	.62	.20	.32	.33	.25
5	.43	.59	.23	.58	.17	.62	.04	.69	.30	.25	.46	.31
$\bar{X}$	.33	.61	.20	.58	.26	.70	.14	.65	.23	.42	.36	.26
s.d.	.10	.04	.13	.02	.09	.10	.05	.04	.05	.14	.08	.08

Directe vergelijking kenmerk- en subkenmerkmethode

Aan de hand van de gegevens uit de extra sessie wordt weer nagegaan of de in tabel 6.13 aanwezige verschillen tussen de beoordelingsmethoden met betrekking tot de overeenstemming met het referentie-oordeel, het gevolg zijn van het verschil in kwaliteit van de beoordelingsmethoden of misschien mede veroorzaakt zijn door de opzet van de training. Als de resultaten van de extra sessie (tabel 6.14 en 6.15) ongeveer hetzelfde beeld laten zien als tabel 6.13, dan worden de verschillen in overeenstemming toegeschreven aan de verschillen tussen de beoordelingsmethoden.

Tabel 6.14: Overeenstemming (Kappa) met het referentie-oordeel per beoordelingsaspect en per beoordelaar (1, 2, 4 en 5) op respectievelijk kenmerk- (K) en subkenmerkniveau (SK). (directe vergelijking)

	beoordelingsaspecten											
	OU		DI		CA		CO		PA		AF	
	K	SK	K	SK	K	SK	K	SK	K	SK	K	SK
1	.50	.66**	.50	.55*	.25	.80**	.25	.91**	-.25	.25	.50	.57*
2	.25	.63*	.00	.90**	.25	.75**	1.**	.82**	.00	.00	.75*	.25
4	.50	.55*	.00	.80**	.75*	.60*	.00	.64*	.00	.00	.50	.57*
5	.50	.55*	-.25	.80**	.00	.65**	.25	.82**	.25	.00	.50	.54*

* = $p<.05$
** = $p<.01$

Op kenmerkniveau zijn slechts drie overeenstemmingscoëfficiënten statistisch significant. In tegenstelling daarmee zijn op subkenmerkniveau de meeste Kappa's significant. De verschillen tussen de beoordelingsmethoden komen duidelijker naar voren in tabel 6.15, die de resultaten van tabel 6.14 samenvat door per beoordelingsaspect de gemiddelde overeenstemming met het referentie-oordeel te geven.
Evenals in tabel 6.13 is de overeenstemming voor "afwerking" groter op kenmerk- dan op subkenmerkniveau en is het verschil tussen beide beoordelingsmethoden, met betrekking tot het beoordelingsaspect "pulpo axiale afschuining", klein. Voor de overige beoordelingsaspecten geldt dat de overeenstemmingen met de referentie-oordelen veel groter zijn als beoordeeld wordt met de subkenmerkmethode. Behalve voor het aspect "afwerking" zijn de standaardafwijkingen aanzienlijk kleiner op subkenmerk- dan op kenmerkniveau. Dit betekent dat de beoordelingsprestaties van de beoordelaars op subkenmerkniveau veel minder van elkaar verschillen dan op kenmerkniveau.
Resumerend kan geconcludeerd worden, dat de resultaten uit de extra sessie bevestigen wat reeds bleek uit de in tabel 6.13 gepresenteerde resultaten van de originele trainings-sessies, namelijk dat de diverse aspecten waarmee klasse II-tweevlakspreparaties gekwalificeerd worden, meer valide beoordeeld worden met de subkenmerk- dan met de kenmerkmethode.

Tabel 6.15: Gemiddelde ($\bar{X}$) overeenstemming (Kappa) met het referentie-oordeel en standaardafwijking (s.d.) op kenmerk-(K) en subkenmerkniveau(SK), berekend per beoordelingsaspect. (directe vergelijking)

	beoordelingsmethode kenmerken		subkenmerken	
beoordelingsaspecten	$\bar{X}$	s.d.	$\bar{X}$	s.d.
outline	.44	.13	.60	.06
diepte	.06	.31	.76	.15
cav.opp.hoek	.31	.31	.70	.09
convergentie	.38	.43	.80	.11
pulpo ax. afsch.	.00	.20	.06	.13
afwerking	.56	.13	.48	.16

6.3.3.3 Discussie

Voor kwalitatief goede beoordelingen is betrouwbaarheid alleen niet voldoende. Beoordelaars kunnen het volstrekt met elkaar eens zijn zonder dat hun beoordelingen representatief zijn voor de kwaliteit van het beoordeelde. Een goed beoordelingsinstrument levert scores op die de kwaliteit van het beoordeelde accuraat beschrijven. Hier dient zich een "kip-en-ei" probleem aan, want wie bepaalt wat de kwaliteit van een werkstuk is? Een echt criterium is er niet; vandaar ook de inspanningen van velen (zie par. 2.3) om het noodzakelijke beoordelen zo objectief mogelijk te laten gebeuren. In het trainingsprogramma is de kwaliteit van de werkstukken vastgelegd in referentie-oordelen. Dit zijn de modale beoordelingen van drie ervaren instructeurs uit het preklinisch onderwijs. De overeenstemming die beoordelaars bereiken met het referentie-oordeel is indicatief voor de accuraatheid waarmee de kwaliteit van het werkstuk of een onderdeel van dat werkstuk beschreven wordt. De resultaten uit de trainings-sessies zijn voor alle beoordelaars hetzelfde:

1. Per werkstuk is de overeenstemming met het referentie-oordeel gemiddeld twee keer zo groot op subkenmerk- als op kenmerkniveau (tabel 6.10). Evenals bij de betrouwbaarheidsschattingen (tabel 6.3) moet ook voor de validiteitsschattingen (tabel 6.10) geconstateerd worden, dat de standaardafwijkingen

bij beide beoordelingsmethoden groot zijn. Ook op subkenmerkniveau worden werkstukken dus regelmatig anders gekwalificeerd dan volgens het referentie-oordeel juist zou zijn.

2. Per beoordelingsaspect bekeken zijn de verschillen tussen de beoordelingsmethoden nog groter. Evenals bij de schattingen voor de betrouwbaarheid (tabel 6.6), moet een uitzondering gemaakt worden voor het beoordelingsaspect "afwerking". Over de kwaliteit van dit aspect zijn de beoordelaars vrij vaak een heel andere mening toegedaan dan die welke beschreven wordt door het referentie-oordeel. Maar onderling zijn de beoordelaars het ook vaak oneens over de kwaliteit van dit aspect. Voor alle andere beoordelingsaspecten blijkt uit vergelijking van tabel 6.6 met 6.13, dat op subkenmerkniveau de bereikte overeenstemmingen met het referentie-oordeel ongeveer even groot zijn als de overeenstemmingen die beoordelaars onderling bereiken. Op kenmerkniveau, daarentegen, bereiken beoordelaars onderling hogere overeenstemmingen dan met het referentie-oordeel voor de meeste beoordelingsaspecten.

Het niet bereiken van overeenstemming met het referentie-oordeel kan een aantal oorzaken hebben, waarvan de voornaamste zijn:
- de beoordelaar faalt;
- de beoordelingsmethode is niet voldoende objectief;
- het referentie-oordeel is niet juist.

Over het falen van beoordelaars en niet-objectieve beoordelingsmethoden is in voorgaande hoofdstukken al uitvoerig gesproken. Indicaties voor het onjuist zijn van de referentie-oordelen zijn de lage overeenstemmingen met het referentie-oordeel, terwijl de beoordelaars het onderling in hoge mate eens zijn. Uitspraken doen over de kwaliteit van de referentie-oordelen op grond van de beoordelingen uit de trainings-sessies, is echter een hachelijke zaak. In de eerste plaats, omdat het verwarrend werkt als de kwaliteit van beoordelingen wordt gemeten via de overeenstemming met het referentie-oordeel en tegelijkertijd de kwaliteit van de referentie-oordelen wordt gemeten via het aantal overeenstemmende beoordelingen. Toch is er geen andere manier om de kwaliteit van de referentie-oordelen te controleren; het is het onvermijdelijke gevolg van het ontbreken van een absolute kwaliteitsbeschrijving. In de tweede plaats, omdat over te weinig waarnemingen beschikt wordt om betrouwbare uitspraken te doen over de kwaliteit van de referentie-oordelen. Voorlopig wordt daarom afgezien van aanpassing van de referentie-oordelen. Eerst als voor elk beoordelingsaspect (op kenmerk- én subkenmerkniveau) van elk werkstuk in het werkstukkenbestand over een groot aantal waarnemingen (beoordelingen van tandarts-instructeurs) beschikt kan worden, kan eventueel tot aanpassing worden overgegaan.

6.3.4 Trainings-effecten

6.3.4.1 Beantwoording van deelvraag F: Is er sprake van een systematisch toenemende beoordelaarsovereenstemming naarmate aan meer trainings-sessies is deelgenomen?

De in de trainings-sessie aangeboden werkstukken waren op toevallige wijze uit het werkstukkenbestand getrokken. Het doel daarvan was de gemiddelde moeilijkheidsgraad van de aangeboden werkstukken in elke sessie ongeveer gelijk te laten zijn. Een belangrijke vraag is of de beoordelingskwaliteit systematisch toeneemt met de hoeveelheid training die beoordelaars ontvangen. De beoordelingskwaliteit wordt afgeleid van de mate van overeenstemming met het referentie-oordeel. Tabel 6.16 geeft per beoordelaar en per training-sessie de berekende overeenstemmingen weer met het referentie-oordeel op respectievelijk kenmerk- en subkenmerkniveau.

Tabel 6.16: Overeenstemming (Kappa) met het referentie-oordeel op kenmerk-(K) en subkenmerkniveau(SK), per beoordelaar (1 t/m 5) en per trainings-sessie (I t/m VI). Onder de stippellijn: gemiddelde overeenstemming ($\bar{X}$) per sessie en standaardafwijking (sd).

	kenmerkmethode						subkenmerkmethode					
	I	II	III	IV	V	VI	I	II	III	IV	V	VI
1	.08	.18	.44	.25	.19	.42	.60	.48	.55	.58	.71	.53
2	.13	.14	.28	.34	.22	.22	.52	.40	.61	.70	.55	.60
3	.33	.25	.19	.44	.19	.25	.57	.40	.57	.69	.54	.49
4	.21	.39	.13	.16	.25	.19	.67	.27	.54	.51	.45	.45
5	.21	.25	.16	.41	.33	.25	.42	.25	.56	.63	.62	.46
$\bar{X}$	.19	.24	.24	.32	.24	.27	.56	.36	.57	.62	.57	.51
sd	.09	.10	.13	.12	.06	.09	.09	.10	.03	.08	.10	.06

Ondanks het feit dat de gemiddelde overeenstemmingen met het referentie-oordeel een stijgende tendens vertonen tot aan de vijfde trainings-sessie, is er op kenmerkniveau geen sprake van

een trainings-effect zoals dit hiervoor gedefinieerd werd. De stijgende tendens is het gevolg van de middeling over de beoordelaars. Als naar de individuele beoordelaar gekeken wordt zijn de resultaten zeer wisselvallig. Kennelijk maakt het voor de beoordelingskwaliteit niet uit of beoordelaars getraind worden of niet. Verwonderlijk is dit niet, gezien de niet-eenduidige criteriaomschrijvingen van de beoordelingsaspecten van de kenmerkmethode. Training heeft geen zin als niet duidelijk is waarop getraind moet worden. Maar ook op subkenmerkniveau is geen sprake van een trainings-effect in de zin van een systematisch toenemende overeenstemming naarmate vaker getraind is. In de eerste sessie worden relatief hoge overeenstemmingen bereikt door beoordelaar 1, 3 en 4. In sessie II relatief lage overeenstemmingen door alle beoordelaars. Van de tweede tot en met de vierde sessie is een opgaande lijn zichtbaar voor beoordelaar 1, 2, 3 en 5. Beoordelaar 1 is de enige die in de vijfde sessie een nog hogere overeenstemming bereikt dan in de vierde sessie.

Tabel 6.17: Inter-beoordelaarsovereenstemming (Kappa) per beoordelaar (1 tot en met 5) en per trainings-sessie (I tot en met VI) op respectievelijk kenmerk- (K) en subkenmerkniveau (SK). Onder de stippellijn: gemiddelde inter-beoordelaarsovereenstemming ($\bar{X}$) per sessie en standaardafwijking (sd)

	kenmerkmethode						subkenmerkmethode					
	I	II	III	IV	V	VI	I	II	III	IV	V	VI
1	.39	.29	.42	.21	.38	.33	.52	.55	.58	.61	.61	.62
2	.30	.38	.38	.38	.30	.31	.45	.52	.65	.69	.58	.58
3	.43	.33	.38	.28	.30	.22	.50	.55	.59	.69	.60	.56
4	.33	.37	.20	.24	.35	.29	.54	.44	.57	.50	.57	.56
5	.35	.43	.35	.35	.40	.30	.45	.52	.64	.60	.60	.56
$\bar{X}$	.36	.36	.35	.29	.35	.29	.49	.52	.61	.62	.59	.58
sd	.05	.05	.09	.07	.05	.04	.04	.05	.04	.08	.02	.03

In tabel 6.16 is beoordelingskwaliteit gedefinieerd als de mate van overeenstemming met het referentie-oordeel. Van een trainings-effect is dan sprake als de overeenstemming met het referentie-oordeel systematisch toeneemt met elke volgende trainingssessie. Als beoordelingskwaliteit niet afgemeten wordt aan een "absoluut" criterium (referentie-oordeel) maar gerelateerd wordt aan de mate van overeenstemming die beoordelaars onderling bereiken, kan opnieuw de aanwezigheid van een trainings-effect onderzocht worden. Tabel 6.17 geeft de resultaten hiervan weer. De overeenstemmingen op kenmerkniveau nemen niet systematisch toe met het aantal trainingen. Wel moet geconstateerd worden dat ze meestal groter zijn dan de overeenstemmingen met het referentie-oordeel in Tabel 6.16. Dit betekent dat op kenmerkniveau de deelnemers aan de trainingen onderling gemakkelijker overeenstemming bereiken over de kwaliteit van werkstukken dan met het referentie-oordeel. Op subkenmerkniveau, daarentegen, zijn er vage aanwijzingen dat een trainings-effect aanwezig is. Tot aan de vijfde trainings-sessie bereiken beoordelaar 1, 2 en 3 een steeds hogere overeenstemming met de andere beoordelaars. Voor beoordelaar 5 is een dergelijke ontwikkeling herkenbaar tot aan de vierde trainings-sessie. Alleen voor beoordelaar 4 is geen enkele systematiek herkenbaar.
De onderlinge overeenstemmingen vertonen een systematischer patroon dan de overeenstemmingen met het referentie-oordeel in tabel 6.16. Maar het lijkt voorbarig om, op grond van bovenstaande resultaten, te concluderen dat er sprake is van een trainingseffect in de zin van hoger wordende overeenstemmingen naarmate vaker getraind is. Een onderzoek dat niet-experimentele variabelen onder controle kan houden is noodzakelijk als zekerheid gewenst is over de vraag of er een positief verband bestaat tussen hoeveelheid training en bereikte overeenstemming. In hoofdstuk VII wordt nader ingegaan op een dergelijk onderzoek.

6.3.4.2 Beantwoording van deelvraag G: Neemt de benodigde beoordelingstijd voor een subkenmerk systematisch af als gevolg van training?

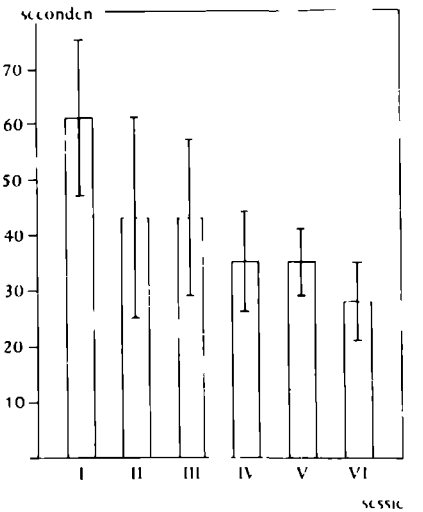
De bruikbaarheid van de subkenmerk-beoordelingsmethode is afhankelijk van de tijd die het werken ermee in beslag neemt. Vandaar dat het belangrijk is om te weten of door training de beoordelingstijd gereduceerd kan worden. In par. 6.2 werd al vermeld dat training geen invloed had op de benodigde tijd voor het beoordelen op kenmerken. Verwonderlijk is dit niet, gezien de langdurige ervaring van de meeste beoordelaars met de kenmerkmethode. Op voorhand wordt verwacht dat training de benodigde tijd voor het beoordelen op subkenmerken kan reduceren. Immers, bij de eerste trainings-sessie worden de beoordelaars geconfronteerd met een beoordelingsmethode die nagenoeg nieuw voor hen is. Alle tekst is nieuw en moet aandachtig gelezen worden om de strekking ervan te kunnen begrijpen. Na verloop van tijd echter, hebben de beoordelaars aan een half woord voldoende om te weten wat er staat. De benodigde beoordelingstijd wordt daardoor korter. Uitsluitsel

over de juistheid van deze verwachting geeft tabel 6.18, waarin per trainings-sessie en per beoordelaar is aangegeven hoeveel seconden gemiddeld besteed werden per subkenmerk-beoordeling.

Tabel 6.18: Gemiddelde beoordelingstijden (in hele seconden) per subkenmerk, per trainings-sessie en per beoordelaar. Gemiddelde beoordelingstijden per subkenmerk en per trainings-sessie ($\bar{T}$) over alle beoordelaars en per beoordelaar ($\bar{B}$) over alle trainings-sessies en standaardafwijking (s.d.).

	trainings-sessies							
beoordelaars	I	II	III	IV	V	VI	$\bar{B}$	s.d.
1	37	21	24	19	25	21	25	7
2	65	48	45	42	35	29	44	12
3	65	42	54	41	37	28	45	13
4	75	69	59	37	41	40	54	16
5	61	36	34	37	35	23	38	13
$\bar{T}$	61	43	43	35	35	28		
s.d.	14	18	14	9	6	7		

Het verband tussen training en gemiddelde beoordelingstijd per subkenmerk wordt duidelijker geïllustreerd in figuur 6.1. Daarin valt duidelijk een dalende tendens te onderscheiden met betrekking tot de hoeveelheid tijd die beoordelaars gemiddeld nodig hebben gehad per subkenmerk. Met name het verschil tussen sessie I en II is opmerkelijk. Deze zeer grote reductie in gemiddelde beoordelingstijd doet zich volgens tabel 6.18 bij bijna elke beoordelaar voor. Kennelijk raakten de beoordelaars erg snel vertrouwd met de inhoud en werkwijze van het beoordelingsprotocol. Na de tweede sessie verloopt de reductie van de benodigde beoordelingstijd per subkenmerk minder snel en is zelfs twee keer sprake van een stabilisatie. Geconcludeerd kan worden, dat de benodigde beoordelingstijd afneemt met een toename van het aantal trainingen. Minstens zo belangrijk is de bevinding in par. 6.3.4.1 dat in elke trainingssessie (behalve de tweede) de overeenstemming met het referentieoordeel ongeveer gelijk bleef. De snellere hantering heeft dus geen nadelige gevolgen voor de beoordelingskwaliteit.



Figuur 6.1: Gemiddelde beoordelingstijd per subkenmerk en standaardafwijking, berekend over alle beoordelaars per trainings-sessie.

Als per beoordelaar de gemiddelde beoordelingstijd per subkenmerk berekend wordt over alle trainings-sessies, dan kan onderscheid gemaakt worden tussen beoordelaars met betrekking tot de snelheid waarmee ze het beoordelingsprotocol hanteren. De verschillen zijn groot (zie tabel 6.18); beoordelaar 1 is meer dan twee keer zo snel als beoordelaar 4. De overige beoordelaars hebben gemiddelde beoordelingstijden die tussen deze twee uitersten liggen.

6.3.4.3 Discussie

De interpretatie van de in par. 6.3.4.1 en 6.3.4.2 gepresenteerde onderzoeksgegevens wordt bemoeilijkt door:
1. het niet consistent zijn van de resultaten in tabel 6.16 en 6.17;
2. het feit dat irrelevante variabelen een rol hebben kunnen spelen.

ad 1. In tabel 6.16 kan geconstateerd worden dat overeenstemmingen met het referentie-oordeel niet systematisch toenemen als gevolg van training. In tabel 6.17 is voor de subkenmerk-beoordelingsmethode wel enig verband herkenbaar tussen overeenstemming en training. Een duidelijke verklaring voor deze inconsistente resultaten kan niet gegeven worden. Wat opgemerkt moet worden is dat de overeenstemmingen in tabel 6.16, berekend over de in de tweede trainings-sessie aangeboden werkstukken (subkenmerkniveau), een opvallende daling vertonen ten opzichte van de eerste sessie. In tabel 6.17 doet dit verschijnsel zich niet voor. Dit betekent dat de beoordelaars het onderling meer eens zijn over de kwaliteit van de in de tweede trainings-sessie beoordeelde

werkstukken, dan dat ze het eens zijn met de door het referentie-oordeel aangegeven kwaliteit. Evenals in par. 6.3.3.3 wordt hier de vraag naar de validiteit van de referentie-oordelen weer actueel. Maar ook nu moet het antwoord weer luiden dat over onvoldoende gegevens beschikt wordt om uitspraken te kunnen doen over de kwaliteit van de referentie-oordelen.

ad 2. In par. 6.3.1 werden al enkele tekortkomingen van de onderzoeksopzet besproken. Het gelijktijdig gebruik van twee beoordelingsmethoden heeft enkele nadelen, waarvan de onvolledige waarnemingen op subkenmerkniveau en de beïnvloeding van de beoordeling op subkenmerkniveau door de teruggekoppelde informatie op kenmerkniveau, de belangrijkste zijn. Andere tekortkomingen betreffen de, onder punt 1 al genoemde, onzekerheid over de kwaliteit van de referentie-oordelen en het feit dat, ondanks de toevallige toewijzing van de werkstukken aan de trainings-sessies, de gemiddelde moeilijkheid van de werkstukken niet gegarandeerd gelijk is geweest in elke trainings-sessie. De conclusie is dat de onderzoeksopzet niet geschikt is voor het opsporen van een verband tussen training en overeenstemming. Het opsporen van een dergelijk verband is alleen mogelijk in een strenger gecontroleerde onderzoeksopzet. Bijvoorbeeld een onderzoek waarin in elke trainings-sessie dezelfde werkstukken (onder een ander identificatienummer) worden aangeboden. Belangrijk is ook dat de beoordelaars allemaal dezelfde aspecten beoordelen en de beoordeling op subkenmerkniveau dus niet afhankelijk is van de beoordeling op kenmerkniveau (zoals in de trainings-sessies gebruikelijk was). Tenslotte zal ook het tijdstip waarop getraind wordt voor alle beoordelaars gelijk moeten zijn. Dit vanwege het te verwachten concentratieverlies dat aan het eind van een werkdag/week eerder zal optreden dan aan het begin.

Trainings-effect is ook geoperationaliseerd als een positief verband tussen de hoeveelheid ontvangen training en de tijd die een beoordelaar gemiddeld nodig heeft gehad voor het beoordelen van een werkstuk op een subkenmerk. Er zijn aanwijzingen voor het bestaan van een dergelijk trainings-effect. Een belangrijke vraag is of dit trainings-effect alléén voldoende rechtvaardiging is voor het invoeren in het onderwijs van trainings-sessies. Verwacht mag worden dat beoordelaars ook zonder training het beoordelingsprotocol zullen leren hanteren; enkel en alleen als gevolg van het gebruik in de praktijk. Training moet om de investering in tijd en geld te rechtvaardigen een "meerwaarde" hebben ten opzichte van "leren-in-de-praktijk". Nagaan of er sprake is van "meerwaarde" vereist een experimentele of quasi-experimentele onderzoeksopzet, waarin de beoordelingsprestaties van een experimentele groep vergeleken kunnen worden met die van een controlegroep. Povenmire & Roscoe (1971) en Roscoe (1971; 1972) stelden een maat voor die de meerwaarde van training kan weergeven: de "Transfer Effectiveness Ratio"(TER). Stel, bijvoorbeeld, dat student-assi-

stenten niet zelfstandig mogen beoordelen voordat is vastgesteld dat zij betrouwbare beoordelaars zijn (minimaal 75 procent inter-beoordelaarsovereenstemming). Onder leiding van een instructeur leren ze beoordelen in de praktijk van het preklinisch motorisch onderwijs. De helft van de groep student-assistenten krijgt daarnaast beoordelaarstraining (experimentele groep). Voor zowel controlegroep als experimentele groep kan de gemiddelde hoeveelheid tijd berekend worden die de student-assistenten nodig hebben gehad om aan het genoemde criterium te voldoen. Stel dat dit voor de controlegroep en experimentele groep respectievelijk 12 en 9 uur bedraagt en dat de experimentele groep in totaal 3 uur training heeft gekregen. De maat TER zou dan bedragen: (12-9)/3 = 1, wat betekent dat één uur training een besparing oplevert van één uur leren beoordelen onder toezicht. Of dit een waarde is die training rechtvaardigt hangt vooral af van de kosten van de training en van het leren beoordelen in de praktijk van het preklinisch motorisch onderwijs. Aan die kostenberekening zitten erg veel, op voorhand onbekende, aspecten. Wat betekent het bijvoorbeeld voor de efficiëntie van het onderwijs als instructeurs minder aandacht hoeven te besteden aan het inwerken van student-assistenten? En hoeveel bedragen de ontwikkelingskosten en operationele kosten van een trainingsprogramma? En beklijven de aangeleerde beoordelings-vaardigheden beter als ze in een trainingsprogramma systematisch zijn geoefend? Deze en nog andere vragen moeten beantwoord worden om precies aan te kunnen geven wat de effectiviteit van training is in de zin van transfer naar een operationele beoordelingssituatie. Belangrijke overwegingen bij de kosten-baten berekening van het onderhavige trainingsprogramma zijn de meervoudige gebruiksmogelijkheid (par. 4.1) en, natuurlijk, de geïndividualiseerde opzet.

6.4 Conclusies

Gelet op de in par. 6.3 gepresenteerde resultaten kunnen de volgende conclusies geformuleerd worden:

1. Beoordelingen van klasse II-tweevlakspreparaties aan de hand van de subkenmerk-beoordelingsmethode zijn betrouwbaarder dan de beoordelingen aan de hand van de kenmerk-beoordelingsmethode.

2. De kwaliteit van vervaardigde klasse II-tweevlakspreparaties wordt accurater beschreven met behulp van subkenmerk-beoordelingen dan met kenmerk-beoordelingen. Anders gezegd: de subkenmerk-beoordelingsmethode is een meer valide beoordelingsinstrument dan de kenmerk-beoordelingsmethode.

3. Trainings-effecten beperken zich hoofdzakelijk tot een systematisch afnemende beoordelingstijd naarmate vaker getraind is.

Naast deze belangrijkste conclusies zijn verder de volgende constateringen nog van belang:

- Anders dan bij de overige beoordelingsaspecten wordt "afwerking" meer valide beoordeeld op kenmerk- dan op subkenmerkniveau.

- Het lijkt zinvol om in toekomstige trainingen extra aandacht te besteden aan het beoordelen van "grensgevallen".

- Het vermoeden bestaat, hoewel gebaseerd op een klein aantal waarnemingen, dat sommige referentie-oordelen van twijfelachtige kwaliteit zijn.

- De betrouwbaarheid van een cijfer gebaseerd op gesommeerde kenmerkscores is vrij laag. Het nemen van zak-/slaagbeslissingen op grond van deze cijfers wordt daarom ontraden (zie bijlage 5). Gezien de grotere betrouwbaarheid van de subkenmerkbeoordelingsmethode is het aan te bevelen om cijfers voor werkstukken te baseren op subkenmerkscores.

- Op basis van de gemiddelde beoordelingstijden in de laatste trainings-sessie kan berekend worden, dat de beoordeling van een werkstuk met de subkenmerkmethode minstens zeven keer zo veel tijd in beslag neemt als met de kenmerkmethode.

VII ALGEMENE DISCUSSIE EN AANBEVELINGEN

Het uiteindelijke doel van de in het eerste deel van deze dissertatie beschreven activiteiten is het verbeteren van de kwaliteit van het preklinisch motorisch onderwijs. De kwaliteit van dit onderwijs wordt door vele factoren beïnvloed. Bijvoorbeeld door de faciliteiten waarover men kan beschikken. Maar ook door de financiële middelen, de studenten, de stafleden, het onderwijsprogramma en het onderzoek. Genoemde factoren echter, zijn veel te globaal om nauwkeurig aan te kunnen geven hoe die beïnvloeding in zijn werk gaat. Het is onmogelijk om het verband tussen, bijvoorbeeld, financiële middelen en de kwaliteit van het onderwijs te beschrijven, zonder in detail in te gaan op wat precies onder beide factoren verstaan moet worden. Spreken over de kwaliteit van onderwijs is spreken over een ingewikkeld raderwerk, waarin elk radertje direct of indirect draait als gevolg van het draaien van andere. Kwalitatief goed onderwijs zal over het algemeen tandartsen afleveren die in staat zijn tot kwalitatief goede dienstverlening. Daaronder wordt verstaan een gevarieerde en excellente dienstverlening. Gevarieerd betekent, dat de tandarts zich laat leiden door feitelijke informatie bij de keuze van een behandeling en zich niet beperkt tot, bijvoorbeeld, restauratie of extractie. Excellente dienstverlening kan vanuit twee invalshoeken bekeken worden: functionaliteit en esthetiek.
Ten einde tandartsen op te kunnen leiden die aan genoemde criteria kunnen voldoen, wordt in het onderwijs veel tijd uitgetrokken voor het aanleren van motorische vaardigheden. Een belangrijke vraag is hoeveel oefening nodig is om van constante, goede kwaliteit verzekerd te zijn. Is de tijd die momenteel in het preklinisch en klinisch onderwijs wordt uitgetrokken voor het aanleren van motorische vaardigheden te ruim, goed of te krap bemeten? Zijn er misschien mogelijkheden om dezelfde doelstellingen in een kortere periode te bereiken? Deze vragen hebben betrekking op de efficiëntie van het leren in de (pre)kliniek.
In par. 1.3.3 werd aandacht besteed aan een viertal door Mackenzie (1973) geopperde benaderingen voor het vergroten van de efficiëntie van het leerproces. Beknopt weergegeven komt het er op neer dat Mackenzie streeft naar het schrappen van overlappingen in de leerstof en naar het construeren van beoordelingsinstrumenten, die niet alleen gebruikt kunnen worden om beslissingen te nemen over studenten, maar ook diagnostisch van nut kunnen zijn. Het onderwijsstimuleringsproject dat in dit deel van de dissertatie beschreven is, heeft in het teken gestaan van de constructie en kwaliteitscontrole van een dergelijk beoordelingsinstrument. Gehoopt wordt dat de inspanningen uiteindelijk zullen leiden tot een efficiënter leerproces. Voorwaarde daarvoor is dat het instrument betrouwbaar en valide is. Het onderzoeken van deze eigenschappen is lastig omdat, evenals de kwaliteit van het onderwijs, de beoordelingskwaliteit afhankelijk is van een groot aantal factoren. De belangrijkste zijn:
1. de beoordelaar;
2. het beoordelingsinstrument;

3. het te beoordelen werkstuk;
4. de beoordelingssituatie.

ad 1. Tenzij er sprake is van een volstrekt objectief beoordelingsinstrument, is de beoordelaar een belangrijke beïnvloedingsbron van de beoordeling. Persoongebonden eigenschappen als genoten opleiding, ervaring met beoordelen en karakter kunnen meespelen in het beoordelingsproces en een onder- of overwaardering van de feitelijke kwaliteit tot gevolg hebben.

ad 2. Beoordelingsinstrumenten kunnen globaal of analytisch zijn, veel of weinig schaalpunten bevatten en met trefwoorden of uitgebreide omschrijvingen werken. De keuze voor een vorm dient afhankelijk te zijn van het doel waarvoor het instrument gebruikt gaat worden. Zoals in hoofdstuk II is beschreven heeft elke keuze gevolgen voor de beoordelingskwaliteit.

ad 3. Ook als beschikt kan worden over een goed beoordelingsinstrument zullen beoordelaars meer moeite hebben met een werkstuk dat net niet of net wel aan de vereisten voldoet dan met een werkstuk dat duidelijk goed of slecht is.

ad 4. De situatie tijdens de beoordeling is de minst voorspelbare factor die de kwaliteit van de beoordeling meebepaalt. Toevallige gebeurtenissen kunnen de beoordeling beïnvloeden. Te denken valt aan plotseling optredend rumoer in de omgeving van de beoordelaar, wisselende licht-omstandigheden of het al dan niet aanwezig zijn van de maker van het werkstuk bij de beoordeling.

Op de meeste van de opgesomde beïnvloedingsfactoren kan niet of nauwelijks invloed worden uitgeoefend. Het beoordelingsinstrument is het voornaamste middel waarmee de beoordelingskwaliteit opgevoerd kan worden. Maar, ongeacht de gedetailleerdheid en objectiviteit van een beoordelingsinstrument kunnen zich toch beoordelingssituaties voordoen waarin de voorschriften van een dergelijk instrument niet voorzien. Het is dan aan de beoordelaar om de voorschriften zodanig te interpreteren dat de beoordeling recht doet aan de objectieve kwaliteit van het beoordeelde. Beoordelaarstraining is er op gericht om beoordelaars te leren hoe prestatiecriteria geïnterpreteerd moeten worden.
In hoofdstuk III en IV is de ontwikkeling beschreven van een nieuw beoordelingsinstrument en van een geïndividualiseerd trainingsprogramma voor beoordelaars en tevens hoe beide middelen geïntegreerd getest werden. De resultaten van deze studie (hoofdstuk VI) geven aanleiding tot een gematigd optimisme. De nieuw ontwikkelde beoordelingsmethode maakt betrouwbaardere en meer valide beoordelingen mogelijk. Verder zijn er aanwijzingen dat de gemiddelde beoordelingstijd per subkenmerk systematisch afneemt als gevolg van training. Maar een duidelijk systematisch verband tussen hoeveelheid training en beoordelaarsovereenstemming werd niet

aangetroffen. Vermeld moet worden, dat de opzet van het onderzoek het optreden van een dergelijk trainings-effect niet bevorderde. Het aanbieden van steeds andere werkstukken in de trainingssessies en het voorwaardelijk gebruik van de subkenmerk-beoordelingsmethode hebben het vergelijken van de beoordelingsprestaties van sessie tot sessie moeilijk gemaakt. Concluderen dat training geen zin heeft zou daarom voorbarig zijn.

De conclusies uit het onderzoek geven aanleiding tot het doen van twee aanbevelingen, die respectievelijk betrekking hebben op de invoering van het beoordelingsprotocol als instrument voor het verwerven en vaststellen van preklinische motorische vaardigheden en op het uitvoeren van een experimenteel onderzoek, dat uitsluitsel moet geven over het rendement van training op de beoordelingskwaliteit en op het leren hanteren van het beoordelingsprotocol. Achtereenvolgens wordt op beide aanbevelingen ingegaan.

Invoering van het beoordelingsprotocol

Om drie redenen is invoering van het beoordelingsprotocol in het preklinisch motorisch onderwijs gewenst:

1. De operationeel gedefinieerde prestatiecriteria vormen een goede leidraad voor het verwervingsproces van de betreffende vaardigheid omdat kennis van de beoordelingscriteria het leerproces "stuurt".
2. Door de grotere betrouwbaarheid van het beoordelingsprotocol in vergelijking met de kenmerk-beoordelingsmethode kunnen studenten met meer vertrouwen afgaan op de beoordeling van hun prestaties en daardoor effectiever leren.
3. Beoordelingen aan de hand van het beoordelingsprotocol zijn informatiever dan beoordelingen aan de hand van de kenmerkmethode omdat beoordelingsscores van eerstgenoemde methode informatie geven over de aard van geconstateerde tekortkomingen. Studenten weten op grond van een beoordelingsscore niet alleen óf ze een fout hebben gemaakt maar tevens welke fout.

Helaas kent het gebruik van het beoordelingsprotocol ook een gevoelig nadeel, namelijk de tijd die nodig is om werkstukken met behulp van dit instrument te beoordelen. Op basis van de informatie in tabel 6.18 kan berekend worden, dat een beoordelaar gemiddeld ongeveer 15 minuten (32 subkenmerken à 28 seconden) nodig heeft voor de beoordeling van een klasse II-tweevlakspreparatie aan de hand van het beoordelingsprotocol. Ter vergelijking: het beoordelen van een zelfde type werkstuk aan de hand van de kenmerkmethode neemt gemiddeld twee minuten tijd in beslag (zie par. 6.2). Dat het beoordelingsprotocol niet zonder meer ingevoerd kan worden in het preklinisch onderwijs mag blijken uit het nu volgende. Volgens de handleiding van het eerstejaars blok "Preparatie en Restauratie" (Instituut Conserverende Tandheelkunde voor Volwassenen, 1983) is in totaal 16 uur uitgetrokken voor het leren vervaardigen van een klasse II-tweevlakspreparatie in een Columbia (kunststof) element. In die 16 uur is tevens de toetspoging en de eerste herkansing opgenomen. Aan studenten wordt geadviseerd om twee tot drie oefenwerkstukken te vervaardigen alvorens een

toetswerkstuk te maken. Uitgaande van 90 studenten, 7 assistenten en 4 te beoordelen werkstukken per student zijn er per assistent ongeveer 770 minuten nodig om alle werkstukken te beoordelen aan de hand van het beoordelingsprotocol. Op een totale beschikbare tijd van 960 minuten per assistent is dat onevenredig veel. Ten einde het beoordelingsprotocol toch in te kunnen voeren in het preklinisch motorisch onderwijs, worden hieronder twee voorstellen gedaan om de tijd die assistenten besteden aan het beoordelen zo beperkt mogelijk te houden. In het eerste voorstel wordt dit bewerkstelligd door een deel van de beoordelingstaak af te stoten naar de studenten; in het tweede voorstel door het beoordelingsprotocol in te krimpen.

voorstel 1: Zelfevaluatie van preklinische practicumwerkstukken

Meer verantwoordelijkheid kan gelegd worden bij de student voor het beoordelen van zijn werkstukken. In de literatuur zijn voorbeelden te vinden van zelfevaluatie in het tandheelkundig onderwijs. In par. 4.1 werden enkele daarvan besproken. Vooralsnog wordt alleen gedacht aan zelfevaluatie van oefenwerkstukken. Het gevaar van zichzelf bevoordelende studenten is dan vrij onwaarschijnlijk. Naast de reeds genoemde voordelen in par. 4.1 is het voor het onderhavige geval natuurlijk het belangrijkste dat docenten meer tijd overhouden voor een intensievere begeleiding van zwakke studenten. Toetswerkstukken worden niet door studenten beoordeeld. Door de omvangrijkheid van de beoordelingstaak echter, kan de student niet meer rekenen op onmiddellijke uitslag. Na het verstrijken van de toetstijd worden de werkstukken ingenomen, gecodeerd door een docent en vervolgens uitgereikt aan studentassistenten die voor de beoordeling zorgdragen. Ter controle van de beoordelingskwaliteit worden sommige werkstukken tevens door een docent beoordeeld. Bij ernstige afwijkingen met de beoordeling van de student-assistent (bijvoorbeeld als Kappa kleiner is dan 0.55) wordt het werkstuk door nog een docent beoordeeld. Als de docenten het eens zijn in hun afwijzing van het oordeel van de student-assistent, dan zullen de beoordelingen van laatstgenoemde vaker aan een controle worden onderworpen.

voorstel 2: Inkrimpen van het beoordelingsprotocol

Een andere mogelijkheid om de benodigde tijd voor beoordeling op subkenmerken te beperken is inkrimping van het beoordelingsprotocol. Hiervoor kunnen twee manieren worden aangegeven:

1. Via een beoordelingsonderzoek wordt uitgezocht welke beoordelingsaspecten de meeste variantie verklaren van de kwaliteit van werkstukken. Beoordelingsaspecten die geen of bijna geen variantie verklaren moeten niet in het beoordelingsprotocol worden opgenomen. Salvendy, et al. (1973) vonden in een dergelijk onderzoek dat het beoordelingsaspect "outline" verreweg de beste predictor was voor de kwaliteit van een klasse I preparatie.
2. Resultaten uit tandheelkundig onderzoek kunnen aanleiding zijn

om te concluderen dat bepaalde beoordelingsaspecten niet zinvol zijn. Volgens Mackenzie (1973) moeten werkstukken alleen beoordeeld worden op aspecten die werkelijk van invloed zijn op de kwaliteit van het product. Hij vraagt zich in dit verband af of het voor de duurzaamheid van een amalgaam-restauratie werkelijk iets uitmaakt of lijnhoeken zijn afgerond of niet.
Beide mogelijkheden om het beoordelingsprotocol zo beknopt mogelijk te maken verdienen aandacht. Dit betekent overigens niet dat het protocol in zijn huidige vorm onbruikbaar zou zijn. Zelfs indien besloten zou worden om de kenmerk-beoordelingsmethode te handhaven, is het verstandig om het beoordelingsprotocol ter beschikking te stellen van studenten. De gedetailleerde prestatiecriteria en de nauwkeurige aanwijzingen voor de beoordeling vormen een goede leidraad voor het efficiënt verwerven van de doelvaardigheden.

Experimenteel vaststellen van het rendement van beoordelaarstraining

In par. 6.3.4.3 werd de aanbeveling gedaan om in een experimenteel onderzoek het effect van training op de beoordelingskwaliteit na te gaan. Voor de opzet van een dergelijk experiment wordt gedacht aan het "post-test only control group design" (Campbell en Stanley, 1971). Dit vanwege de eenvoud (geen pretest) en vanwege het feit dat dit model voldoet aan de eisen die nodig zijn om interne validiteit van het experiment te bewerkstelligen. Interne validiteit wil zeggen dat eventuele verschillen tussen groepen alleen veroorzaakt worden door de experimentele conditie. Interne validiteit vereist controle van buitenexperimentele variabelen zoals rijping van de proefpersonen, testeffecten, enz. In het experiment worden proefpersonen (student-assistenten) per toeval toegewezen aan de experimentele en de controle-groep. In de experimentele conditie worden de proefpersonen getraind in het beoordelen door hen terugkoppeling te verstrekken over de geleverde beoordelingsprestaties. Tevens leren deze proefpersonen beoordelen onder leiding van een instructeur in de praktijk van het preklinisch motorisch onderwijs. Elke instructeur participeert in een calibratietraining, voorafgaand aan het experiment, totdat minstens 75 procent overeenstemming wordt bereikt met alle andere instructeurs met betrekking tot een bepaalde beoordelingstaak. Proefpersonen in de controlegroep krijgen geen training. Bij het leren beoordelen in de praktijk dient er voor gezorgd te worden dat alle student-assistenten evenveel werkstukken te beoordelen krijgen. Voor alle proefpersonen wordt vastgesteld hoe lang het duurt voordat zij gemiddeld 75 procent overeenstemming bereiken met de instructeur, onder wiens leiding zij beoordelen. Als het criterium bereikt is worden in een nameting een aantal werkstukken (identiek voor elke student-assistent) ter beoordeling aangeboden. Het beoordelen in de trainings-situatie, in de praktijksituatie en tijdens de nameting dient steeds gestandaardiseerd te verlopen. Dat wil zeggen dat het werkstuk in de fantoomkop gemonteerd moet zijn en dat de beoordelingspositie voorgeschreven is.
Deze onderzoeksopzet maakt het mogelijk om de in par. 6.3.4.3

genoemde meerwaarde van training uit te drukken in een tijd-ratio en daarnaast vast te stellen of training resulteert in kwalitatief betere beoordelingen. Significante verschillen in beoordelingsprestaties tussen de experimentele en controlegroep kunnen worden toegeschreven aan de invloed van de beoordelaarstraining. Bij het uitblijven van significante verschillen moet worden afgezien van het invoeren van beoordelaarstraining als middel om de beoordelingskwaliteit, gemeten via de inter-beoordelaarsovereenstemming en de overeenstemming met het referentie-oordeel, te vergroten. Niettemin verdient het in dat geval aanbeveling tot invoering over te gaan van een éénmalige (jaarlijkse) training, om nieuwe beoordelaars snel vertrouwd te maken met de beoordelingsmethode en met de belangrijkste aspecten van de aan te leren vaardigheden.

DEEL II

METEN VAN PROBLEEMOPLOSVAARDIGHEID

I PROBLEEMOPLOSSEN EN PROBLEEMOPLOSVAARDIGHEID

1.1 Inleiding

Onze wereld is vol problematiek. Het lezen van een krant of het kijken naar het televisiejournaal zijn activiteiten die onmiddellijk leiden tot confrontatie met problemen van de meest uiteenlopende aard. Zo is er in Nederland onder andere sprake van een werkloosheidsprobleem, een energieprobleem, een vervuilingsprobleem, een vergrijzingsprobleem, een automatiseringsprobleem, een verzuilingsprobleem en (vooral) een geldprobleem. Met deze opsomming zijn lang niet alle door de Nederlandse bevolking als probleem gepercipieerde moeilijkheden genoemd. Ook hoeft het niet zo te zijn dat elke Nederlander de opgesomde problemen persoonlijk als probleem ervaart. Een egocentrisch ingestelde werknemer hoeft een leger van één miljoen werklozen niet per se als problematisch te ervaren. Het oplossen van een vergelijking met één onbekende is doorgaans geen probleem voor abituriënten met wiskunde I in het examenpakket. Het begrip "probleem" is dus subject-gebonden, hetgeen tot uiting komt in de definitie zoals Frijda en Elshout (1976) die geven: "Een probleem kan gedefinieerd worden als een situatie waarin het subject is geconfronteerd met een taak, opgave of moeilijkheid waarop hij geen onmiddellijk antwoord weet, en waarop hij ook niet door middel van een geautomatiseerde reeks handelingen een antwoord kan vinden". Deze definitie impliceert dat een leerproces heeft plaatsgevonden zodra de oplossing voor het probleem gevonden is. Met name de "schematheorie" biedt een aannemelijke verklaring voor het leren in een probleemsituatie. Een schema is een cognitieve structuur met betrekking tot een bepaald aspect van de werkelijkheid. Die structuur is aanwezig in het lange-termijn geheugen en bestaat uit abstracte kennis over wat een groot aantal situaties, gebeurtenissen of dingen met elkaar gemeen hebben (Anderson en Pichert, 1978 geciteerd in Schmidt, 1982). Als binnenkomende informatie "past" in reeds aanwezige schema's, dan wordt die informatie daarin opgenomen, zonder dat de structuur van het schema zich wijzigt. Dit proces heet assimilatie. Past de nieuwe informatie niet in het bestaande schema dan accomodeert het schema aan de nieuwe informatie. Op deze wijze worden de schema's steeds uitgebreider en kunnen nieuwe problemen sneller worden opgelost. Leren, zoals het verklaard wordt met de schematheorie, en probleemgeoriënteerd onderwijs sluiten goed op elkaar aan. Volgens de schematheorie komt een leerproces namelijk gemakkelijker op gang als:

- reeds verworven kennis geactiveerd wordt;
- de situatie waarin geleerd wordt zoveel mogelijk lijkt op de situatie waarin de kennis gebruikt moet gaan worden;
- gelegenheid wordt geboden om de verworven kennis te bewerken (Schmidt, 1983).

Wat probleemgeoriënteerd onderwijs is, waarom het momenteel zo veel belangstelling geniet en wat de ervaringen in Nederland zijn met deze vorm van onderwijs, komt ter sprake in par. 1.2. De

uiteindelijke doelstelling van probleemgeoriënteerd onderwijs is het opleiden tot "probleemoplosvaardigheid". Wat dit behelst en hoe het "gemeten" kan worden komt aan de orde in par. 1.3 respectievelijk par. 1.5. In par. 1.4 wordt het toepassen van heuristische methoden bij het oplossen van problemen besproken. Het hoofdstuk wordt afgesloten met een discussie in par. 1.6.

1.2 Probleemgeoriënteerd onderwijs

De fundamenten voor probleemgeoriënteerd onderwijs zijn voor een belangrijk deel gelegd door Jerome Bruner. Deze cognitieve psycholoog heeft grote bekendheid gekregen door de onderwijskundige uitwerking van zijn ideeën over het verwerven van kennis door mensen. Met name geldt dit zijn pleidooi voor "discovery-learning". Volgens Bruner (1961, geciteerd in Schmidt, 1982) heeft "ontdekkend leren" de volgende voordelen boven het traditionele leren, waarbij de activiteiten van de leerlingen voornamelijk beperkt blijven tot passieve kennisverwerving:
- ontdekkend leren leidt tot betere retentie van het geleerde;
- ontdekkend leren bevordert de transfer van het geleerde;
- ontdekkend leren bevordert de intrinsieke motivatie tot leren;
- middels ontdekkend leren oefenen leerlingen het gebruik van heuristische onderzoeksstrategieën.

Bruner's verwachtingen konden niet door empirisch onderzoek bevestigd worden. Met betrekking tot het verrichte onderzoek op dit gebied spreekt van Oers (1981) over "....een ondoorzichtige, warrige collage van empirische resultaten. Van een duidelijke tendentie die houvast biedt voor verder onderzoek of voor de inrichting van onderwijsleersituaties, is geen sprake."
In een artikel over de cognitieve effecten van probleemgestuurd onderwijs haalt Schmidt (1982) Mayer aan, die de teleurstellende resultaten van het "ontdekkend leren" verklaart uit het feit dat het slechts leidt tot activatie van bestaande kennis. Als daarop geen confrontatie met nieuwe kennis volgt, kan geen leren plaatsvinden. Volgens Bruner moeten de lerenden zélf de noodzakelijke nieuwe informatie produceren. Aan de haalbaarheid daarvan moet ernstig getwijfeld worden. Met name op dit laatstgenoemde standpunt van Bruner onderscheidt het probleemgeoriënteerd onderwijs zich van het ontdekkend leren. Bij probleemgeoriënteerd onderwijs moeten de lerenden de hen voorgelegde problemen met meer of minder hulp zien op te lossen.
Schmidt (1979) maakt onderscheid tussen probleemgeoriënteerd onderwijs en probleemgestuurd onderwijs. Probleemgeoriënteerd onderwijs wordt door hem gehanteerd als een overkoepelende term voor een verscheidenheid aan onderwijsvormen die alle gekenmerkt worden door het feit dat studenten aan problemen of taken werken. De term op zich zegt niets over de vorm van het onderwijs, die in principe kan variëren van hoorcollege tot geprogrammeerde instructie. Met probleemgestuurd onderwijs doelt Schmidt op onderwijs waarbij studenten kennis verwerven met behulp van de probleemoplossende methode. Het leerproces vangt aan door studenten te confronteren met een probleem, dat ze met de huidige kennis die

ze bezitten niet adequaat kunnen oplossen. Schmidt (1979): "Het probleem of de taak is daar dus niet het terrein waarop je je reeds verworven kennis en vaardigheden toepast, maar juist een stimulans waardoor het leerproces in gang gezet wordt". Probleemgestuurd onderwijs is in deze visie dus een methode van onderwijs. In een ander artikel zegt Schmidt (1982) hierover: "Probleemgestuurd onderwijs moet in het licht van de schematheorie gezien worden als een elaboratietechniek. Het denken en discussiëren over het voorgelegde probleem activeert bestaande schema's die meer of minder relevant kunnen zijn ten aanzien van dat probleem. Gebaseerd op deze probleemrelevante schema's zullen inferenties (hypothesen) geproduceerd worden waarmee de studenten proberen een eigen cognitieve representatie op te bouwen van de processen die verantwoordelijk geacht kunnen worden voor de in het probleem beschreven verschijnselen. Als het probleem didactisch goed gekozen is, dat wil zeggen als het met bestaande kennis niet bevredigend kan worden opgelost, zal die cognitieve representatie een nieuwe, meer gedifferentieerde constructie zijn, gebaseerd op de gezamenlijke voorkennis van de groepsleden (Schmidt gaat uit van probleemgeoriënteerd groepsonderwijs), en in die zin een herstructurering daarvan".

In navolging van Schmidt zal in het vervolg van dit proefschrift de term "probleemgeoriënteerd onderwijs" gebruikt worden als aanduiding voor alle onderwijssituaties waarin problemen worden opgelost en "probleemgestuurd onderwijs" als dié specifieke vorm van probleemgeoriënteerd onderwijs, waarin studenten kennis verwerven met behulp van de probleemoplossende methode.

In Nederland komt het probleemgeoriënteerd onderwijs steeds meer onder de aandacht van de docenten uit (vooral) het Hoger Onderwijs. Dit staat wellicht in verband met de constatering van menig opleider dat studenten, na beëindiging van een cursus of opleiding, de leerstof wel op reproductieniveau beheersen maar geen zinnig gebruik weten te maken van die stof in probleemsituaties. Willems (1978) wijt dit aan het feit dat de verworven kennis niet instrumenteel is, waardoor die alleen in de context van de bijbehorende informatie oproepbaar is. Ook het feit dat studenten niet beschikken over operaties om de opgeslagen kennis mee toe te passen, speelt hierbij een rol. Probleemgeoriënteerd onderwijs kan deze manco's opheffen omdat de kennis wordt aangeboden in het kader van het werken aan problemen. Dit wordt bevestigd door resultaten uit enkele evaluatiestudies met betrekking tot probleemgeoriënteerd onderwijs. Uit die studies is gebleken dat:

1. studenten over het algemeen enthousiast zijn over de andere inrichting van dit onderwijs (Crombag, 1973; Vaags, 1975; Mettes en Pilot, 1980; de Graaff et al., 1982);
2. prestaties over het algemeen hoger zijn als studenten probleemgeoriënteerd onderwijs hebben genoten (Crombag, 1973 ; van den Briel van Ingen en Plasschaert, 1977; Mettes en Pilot, 1980; Schmidt, 1982; Boshuizen en Claessen, 1982).

Ten aanzien van de eerste conclusie is het verstandig om zich te realiseren dat studenten snel enthousiast zijn als de sleur doorbroken wordt. Het kan echter ook voorkomen dat studenten dermate vertrouwd zijn met een bepaalde manier van onderwijs, dat

ze zich negatief uitlaten over pogingen om die te wijzigen (van den Briel van Ingen en Plasschaert, 1977). Over de tweede conclusie kan gezegd worden dat de gunstige resultaten mede veroorzaakt kunnen zijn door de veelal positieve houding van de studenten ten opzichte van de experimenten (Hawthorne-effect).

1.3 Probleemoplosvaardigheid

Een belangrijk aspect van probleemgeoriënteerd onderwijs is het vaststellen van probleemoplosvaardigheid. Het is verleidelijk om dit begrip te definiëren als de bekwaamheid om problemen op te lossen. In de eerste plaats, echter, is deze definitie veel te ruim, omdat over problemen in het algemeen gesproken wordt. De laatste jaren zijn steeds meer onderzoekers tot de conclusie gekomen dat het onderscheid tussen probleemoplosvaardigheid en kennis niet zo scherp is als altijd werd aangenomen (Berner et al., 1977; Greeno, 1980). Probleemoplosvaardigheid heeft specifiek betrekking op het vakgebied waarin die vaardigheid aangewend wordt. In de tweede plaats kan deze definitie te gemakkelijk leiden tot de opvatting dat een oplossing van het probleem de aanwezigheid van probleemoplosvaardigheid aantoont. Dat dit niet zo is wordt geïllustreerd met volgende voorbeelden:

1. Een vijfdejaars leerling van het Atheneum weet na vijftien minuten de juiste oplossing te geven voor een wiskundige vergelijking met één onbekende.

Het belangrijkste argument om te bestrijden dat deze leerling probleemoplosvaardigheid bezit in het genoemde domein, ligt in de hoeveelheid tijd die het "oplossen" in beslag heeft genomen. De meest voor de hand liggende verklaring voor het feit dat de leerling zo veel tijd nodig heeft gehad, is dat niet over de benodigde kennis beschikt werd om dergelijke problemen op te lossen en daarom werd overgegaan tot "proberen". Duidelijk zal zijn dat hier geen sprake is van probleemoplosvaardigheid met betrekking tot genoemd domein.

2. Een huisarts benadert het maagpijn-probleem van zijn patiënt met het voorschrijven van medicijnen.

Uit het feit dat het probleem voor de patiënt is opgelost kan niet zonder meer worden afgeleid dat de arts probleemoplosvaardigheid bezit. Daarvoor is informatie nodig over het proces dat voorafging aan het voorschrijfgedrag. Als het hulpverleningsgedrag van de arts zich beperkt heeft tot het voorschrijven van medicijnen, dan zijn twijfels over zijn probleemoplosvaardigheid gerechtvaardigd. Als hij, daarentegen, geprobeerd heeft om de oorzaken van de klacht te achterhalen, dan kan het voorschrijven van medicijnen een weloverwogen keuze zijn geweest en een goede oplossing voor het probleem.

3. De spoorwegen hebben hun vervoerscapaciteits-probleem op de lijn Amsterdam-Utrecht opgelost door de aanleg van een nieuwe spoorlijn.

Ook nu is het probleem opgelost, maar of de verantwoordelijke besluitvormers gekwalificeerd kunnen worden als "probleemoplosvaardig" is niet duidelijk. Daarvoor is informatie nodig over de kosten-baten berekening van de gekozen oplossing en mogelijke alternatieve oplossingen, zoals bijvoorbeeld de aanschaf van dubbeldek-wagons. Daarnaast is het belangrijk om te weten of de nieuwe spoorlijn niet te veel geluidsoverlast geeft en/of een bedreiging kan vormen voor bepaalde natuurgebieden. En natuurlijk mag de werkgelegenheid bij de keuze voor een der oplossingen niet uit het oog worden verloren.

Genoemde voorbeelden illustreren dat het niet mogelijk is om te concluderen dat een persoon probleemoplosvaardigheid bezit, enkel en alleen op grond van de constatering dat het voorgelegde probleem door hem werd opgelost. Informatie over besluitvormingsprocessen, gehanteerde probleemaanpak en benodigde tijd is nodig om het feit van de bereikte oplossing naar waarde te kunnen schatten.
Marshall (1983) geeft een algemeen toepasbare definitie van probleemoplosvaardigheid: "Competence in problem solving is measured by the degree of succes achieved in providing a satisfactory solution to undifferentiated presenting situations using a standard considered adequate for the discipline in an economy of time, at minimal expense and causing the least inconvenience". Deze definitie geeft aan dat veel aandacht besteed moet worden aan de wijze waarop een probleem wordt opgelost. Lang niet alle oplossingen voor een probleem zijn acceptabel omdat ze of te veel tijd kosten, te veel geld of te veel ongemak veroorzaken. Wat te veel tijd, geld of ongemak is kan alleen in een concrete probleemsituatie duidelijk worden. Een wat vagere eis is wél in zijn algemeenheid bespreekbaar, namelijk efficiëntie. Het oplosproces is efficiënt als gebruik gemaakt wordt van procedures die de oplossing op systematische wijze dichterbij brengen. In par. 1.4 worden enkele van deze procedures besproken.

1.4 Procedures voor efficiënt probleemoplossen

"Efficiënt probleemoplossen bestaat bij de gratie van efficiënte methodes van exploratie van de zoekruimte, en de theorie van het probleemoplossen is voor een groot deel de theorie van efficiënte exploratiemethodes in doorgaans zeer immense zoekruimtes" (Frijda en Elshout, 1976). Deze uitspraak kan alleen begrepen worden in het kader van de "informatie-verwerkingsbenadering" van het probleemoplossen, zoals die door Newell en Simon (1972) beschreven is. Daarom wordt nu beknopt een beschrijving gegeven van deze benadering van menselijke informatieverwerking. De beschrijving is overgenomen uit Frijda en Elshout (1976).

Probleemoplossen wordt opgevat als een activiteit waarbij getracht wordt om een bepaalde doeltoestand te realiseren. Een "doeltoestand" is een toestand van het probleemmateriaal die met de doelstelling van de probleemoplosser overeenkomt. Een voorbeeld van een doeltoestand is de matstelling bij het schaakspel. Behalve een doeltoestand is er ook een uitgangssituatie, de initiële toestand. Doeltoestand en initiële toestand zijn allebei probleemtoestanden. Een probleemtoestand is iedere toestand van het probleemmateriaal die door beschikbare en toelaatbare operaties uit de initiële toestand ontstaan kan zijn. Een operatie is een transformatieregel die een probleemtoestand kan doen overgaan in een andere. Als operaties op een systematische wijze georganiseerd worden ten einde een doeltoestand te bereiken, dan is er sprake van een methode. De taak van een persoon bij het probleemoplossen bestaat eruit, zodanige operaties of operatieopeenvolgingen te vinden, die toegepast op de initiële toestand, een doeltoestand opleveren. Een doeltoestand moet worden gevonden door operaties toe te passen op een initiële toestand en op de produkten van eerdere operatietoepassing. Maar over het algemeen kunnen op een probleemtoestand meerdere operaties worden toegepast; er zijn, in een bepaalde situatie, doorgaans meerdere dingen die men zou kunnen doen. Probleemoplossen speelt zich dus af in een "ruimte" van mogelijke probleemtoestanden. De paden in die ruimte worden gevormd door de operaties die de ene toestand in de andere doen overgaan. De taak bij het probleemoplossen is om in die ruimte, uitgaand van de initiële toestand, het pad naar de doeltoestand te vinden. In dit verband spreekt men van de "zoekruimte" of "probleemruimte".
Newell en Simon beschouwen het probleemoplossingsproces dus als een zoekproces door de probleemruimte. Het exploreren van deze probleemruimte kan op drie manieren geschieden, namelijk op algorithmische, blinde en heuristische wijze. Van algorithmische exploratie is sprake als er een voorschrift bestaat dat de oplosser in staat stelt om bij ieder knooppunt (probleemtoestand) in de zoekruimte één - de juiste of de beste - operatie te kiezen. Bij blinde exploratie worden één of meerdere operaties onderzocht. De keuze voor die operaties is echter niet gemotiveerd. Als er heuristisch geëxploreerd wordt dan worden op elk knooppunt in de probleemruimte slechts enkele mogelijkheden onderzocht, en wel dié welke plausibel zijn in verband met het vinden van een doeltoestand. De keuze van plausibele handelwijzen wordt gemaakt op grond van niet-lokale informatie, dat wil zeggen, informatie die wordt afgeleid uit grotere gedeelten van de probleemruimte dan het betreffende knooppunt alleen. De methoden die de keuze van plausibele voortzetting in de zoekruimte bepalen noemt men "heuristische methoden". Toepassing van heuristische methoden is noodzakelijk wanneer, voor het probleemgebied in kwestie, een algorithme ontbreekt en de potentiële zoekruimte te groot is om in de beschikbare tijd, met de beschikbare inspanning of met de beschikbare geheugenruimte, systematisch te worden onderzocht. Dat het hier niet gaat om algemeen toepasbare oplossingsmethoden maken Frijda en Elshout (1976) duidelijk door te benadrukken dat toepassing van heuristische methoden alleen mogelijk is als de

probleemruimte niet-lokale informatie bevat. Dat wil zeggen wanneer de probleemoplosser iets van de situatie begrijpt of weet.

Hayes (1981) deelt heuristische procedures in vier algemene klassen in:
1. trial-and-error;
2. nabijheids-methoden;
3. fractionerings-methoden;
4. kennis-gebaseerde methoden.

ad.1 trial-and-error

In deze klasse vallen twee procedures waarmee de probleemruimte afgezocht kan worden:
- blind zoeken;
- systematisch zoeken.

Hayes vat, in tegenstelling tot Frijda en Elshout (1976), "blind zoeken" en "systematisch zoeken" op als heuristische procedures. Laatstgenoemde auteurs vatten deze procedures samen onder de term "blinde exploratie". Blind zoeken wil zeggen dat de oplosser geen gebruik maakt van informatie om bepaalde mogelijkheden te onderzoeken en ook niet bijhoudt of bepaalde mogelijkheden al onderzocht zijn. Als er systematisch gezocht wordt houden oplossers bij welke mogelijkheden ze al onderzocht hebben en kiezen alleen nog niet onderzochte mogelijkheden. Trial-and-error methoden zijn alleen bruikbaar als de probleemruimte niet te groot is; dat wil zeggen, als het pad dat naar de oplossing leidt er niet een is temidden van zeer veel andere paden. Zo zijn de trial-and-error methoden bijvoorbeeld niet geschikt om de combinatie te achterhalen van een cijferslot, omdat het aantal combinaties toeneemt als een functie van het aantal palletjes en het aantal standen van elk palletje. Stel dat het aantal palletjes (p) gelijk is aan 5 en dat elk palletje 10 standen (s) heeft. Het aantal combinaties is dan gelijk aan $s^p = 10^5 = 100000$. Duidelijk is dat trial-and-error methoden voor dergelijke problemen niet geschikt zijn.

ad.2 nabijheidsmethoden

Nabijheids-methoden verschillen van trial-and-error methoden in het feit dat ze één stap vooruit kijken. De belangrijkste vraag bij deze heuristische methoden is: welke volgende stap kan ik ondernemen om dichter bij het doel te komen? Geen enkele nabijheids-methode kijkt naar de moeilijkheden die eventueel optreden nadat de stap gezet is. Bekende nabijheids-methoden zijn:
- de hill-climbing methode;
- middel-doel analyse.

De hill-climbing methode heeft zijn naam te danken aan een verdwaalde, die in een donkere nacht besluit om een heuvel te beklimmen ten einde tekens te kunnen bespeuren die zouden kunnen wijzen op de nabijheid van de bewoonde wereld. In zijn algemeenheid is de methode bruikbaar als de terugkoppeling die verkregen wordt over de ondernomen stap kan leiden tot een volgende stap. Hayes (1981) geeft het voorbeeld van iemand die een televisietoestel

opnieuw moet afstellen. Er wordt een knop gekozen en gedraaid terwijl het scherm aandachtig in de gaten wordt gehouden. Bij een kwaliteitsverbetering van het beeld wordt verder gedraaid in dezelfde richting. Bij een verslechtering van de beeldkwaliteit wordt de tegenovergestelde richting geprobeerd. Als het beeld slechter wordt onafhankelijk van de richting waarin gedraaid wordt, dan wordt een andere knop gekozen. Omdat de afstelling van de ene knop de afstelling van de andere kan beïnvloeden, is het niet uitgesloten dat reeds afgestelde knoppen opnieuw afgesteld moeten worden.
Middel-doel-analyse (Newell en Simon, 1972) verschilt van de hill-climbing methode op de volgende punten:
- middel-doel analyse kan meerdere verschillen tussen initiële toestand en doeltoestand verwerken;
- middel-doel analyse leidt tot formulering van subdoelen waardoor de volgende te ondernemen stap eenvoudiger te bepalen is.

Middel-doel analyse bestaat uit de volgende procedures:
1. Er wordt een lijst opgesteld van verschillen tussen initiële toestand en doeltoestand.
2. Voor het eerste verschil wordt een operatie gezocht die geschikt is om het verschil weg te nemen. Als geen geschikte operatie gevonden kan worden, wordt het volgende verschil op de lijst genomen.
3. Nagegaan wordt of aan de voorwaarden voor toepassing van de operatie voldaan is. Indien dat niet het geval is moet geprobeerd worden om eventuele belemmeringen te elimineren.

Nabijheids-methoden zijn, doordat ze slechts één stap vooruit kijken, ongeschikt om toegepast te worden op problemen waarin op andere dan rechtstreekse wijze naar de doeltoestand wordt gewerkt. Het schaakspel waarin een stuk wordt geofferd om enkele zetten later een strategisch betere positie te kunnen innemen, is een goed voorbeeld van het type probleem ("omweg-probleem") dat niet kan worden opgelost door gebruik te maken van nabijheids-methoden.

ad.3 fractioneringsmethoden

Onder fractionering verstaan Newell en Simon (1972) het opsplitsen van een probleem in deelproblemen, waarbij die deelproblemen ieder voor zich eenvoudiger oplosbaar zijn. Hayes (1981) spreekt niet over deelproblemen maar over subdoelen. Hij noemt als belangrijkste voordelen van fractionering van problemen:
- subdoelen maken het oplossen van problemen gemakkelijker omdat ze de zoekruimte reduceren;
- subdoelen zijn een leidraad voor de oplosser in omweg-problemen.

In de meeste gevallen echter, krijgt de probleemoplosser de subdoelen niet aangereikt; hij moet die zelf zien te vinden. Als de oplosser bekend is met het soort probleem dat voorgelegd wordt, dan worden subdoelen veelal gegenereerd op basis van ervaring. Zo begint bijna iedereen die een meetkundig probleem moet oplossen met het tekenen van een figuur. Bij minder bekende problemen kunnen subdoelen geïdentificeerd worden door het analyseren van de doeltoestand of door gewoon te beginnen.

Volgens Frijda en Elshout (1976) is fractionering mogelijk en zinvol als heuristische methode, wanneer de oplossing van een deelprobleem geëvalueerd kan worden; dat wil zeggen, wanneer vastgesteld kan worden of men met het oplossen van een subprobleem dichter bij het uiteindelijke doel gekomen is.

ad.4 kennis-gebaseerde methoden

Onderscheid kan gemaakt worden in een methode die gebruik maakt van reeds verworven kennis en een methode die nieuwe kennis verwerft met het doel om het probleem op te lossen. De eerste methode berust voornamelijk op het herkennen van relevante overeenkomsten tussen de probleemsituatie en soortgelijke probleemsituaties die in het geheugen zijn opgeslagen (pattern matching). Laatstgenoemde methode houdt in dat het gepercipieerde gebrek aan kennis wordt opgeheven doordat de probleemoplosser gaat lezen over het probleem, of er met deskundigen over gaat praten. Een andere mogelijkheid is dat de oplosser een hulpprobleem construeert; een vereenvoudigde versie van het voorgelegde probleem. Door het hulpprobleem op te lossen kan inzicht verworven worden over hoe het meer ingewikkelde probleem opgelost kan worden.

Voorgaande beschrijving beoogt aan te geven dat de aard van een probleem beperkingen oplegt aan de keuze van een heuristische oplosmethode. De volgende beperkingen werden genoemd:

1. trial-and-error methoden zijn ongeschikt voor toepassing op problemen met een grote zoekruimte;
2. nabijheids-methoden zijn ongeschikt voor problemen die via een "omweg" opgelost moeten worden;
3. fractionerings-methoden zijn ongeschikt als geen terugkoppeling verkregen kan worden over de zinvolheid van de deeloplossing voor de oplossing van het totale probleem.

Om het probleemoplossen efficiënt te laten verlopen is het niet voldoende om gebruik te maken van heuristische methoden. Het is belangrijker dat een oplosser de juiste heuristische methode selecteert op grond van de kenmerken van het probleem en de eisen die aan de oplossing gesteld worden. Als bij de keuze ook nog aspecten meespelen als "gevolgen van falen" en "beperkte tijd", dan wordt wel gesproken van een "strategie" (Bruner, Goodnow en Austin, 1956, geciteerd in Frijda en Elshout, 1976). Een strategie kan gezien worden als een algemeen actieplan waarin is vastgelegd in welke volgorde stappen in het oplosproces moeten plaatsvinden (De Jong en Ferguson-Hessler, 1984). Een bekende Nederlandse studie over strategiegebruik bij het oplossen van problemen is het onderzoek van Mettes en Pilot (1980), waarin geëxperimenteerd werd met een zogenaamd "gewenst handelingsverloop" voor het oplossen van thermo-dynamische problemen. Een groot probleem bij dergelijke strategie-studies is het vaststellen van de effecten die het gebruik van een strategie heeft op de kwaliteit van de oplossing. Aan de oplossing zelf kan niet afgelezen worden of de strategie geheel of slechts gedeeltelijk werd toegepast. Mettes en Pilot (1980) ontwikkelden voor dit doel een speciaal werkblad. De Jong en Ferguson-Hessler (1984) lieten proefpersonen problemen

hardop denkend oplossen en maakten daarvan bandopnames. In het medisch onderwijs wordt veel gewerkt met "papieren simulatie" om het oplosproces van studenten vast te leggen. In par. 1.5 wordt uitvoerig op laatstgenoemde methode ingegaan.

1.5 Het vaststellen van probleemoplosvaardigheid door middel van papieren simulatie

1.5.1 Inleiding

In het geneeskunde- en tandheelkunde-onderwijs wordt veelvuldig gebruik gemaakt van simulatie-technieken om klinische vaardigheden aan te leren en te evalueren. Voor de hand liggende argumenten zijn dat het oefenen op echte patiënten niet altijd zonder risico is en dat niet altijd over voldoende patiënten beschikt kan worden om alle studenten in voldoende mate te kunnen laten oefenen. Het oefenen van motorische vaardigheden geschiedt met behulp van "fantomen", substituten van menselijke organen of ledematen. Tandheelkunde studenten leren "gaatjes boren" in plastic elementen die in een fantoomkop geplaatst zijn. De Medische Faculteit van de Rijksuniversiteit Limburg te Maastricht beschikt over een zogenaamd "Skillslab", waar studenten vanaf het eerste studiejaar kunnen oefenen op fantomen om basisvaardigheden aan te leren. Echter, voorafgaand aan het uitvoeren van handelingen ter oplossing van een geneeskundig of tandheelkundig probleem moeten artsen en tandartsen kunnen beschikken over zogenaamde diagnostische vaardigheden. In het diagnostisch proces spelen communicatieve en probleemoplossende vaardigheden een belangrijke rol, zodat het oefenen op echte patiënten hier onontbeerlijk lijkt. Maar, door de snel groeiende aantallen studenten ontstond in de jaren zestig een tekort aan "geschikte" patiënten, waardoor het noodzakelijk werd om gebruik te maken van simulatiepatiënten. Simulatiepatiënten zijn gezonde mensen die de rol spelen van een patiënt. De grootste betekenis ligt bij het anamnestisch gedeelte van het arts-patiëntcontact. Lichamelijk onderzoek stuit vaak op moeilijkheden omdat het onmogelijk is om alle, bij een bepaald ziektebeeld behorende symptomen te simuleren (Metz, 1984). Een ander probleem dat zich voordoet bij de inschakeling van simulatiepatiënten betreft de standaardisatie. Wil men de prestaties van studenten op examens kunnen vergelijken, dan is het gewenst om alle studenten hetzelfde "probleem" aan te bieden. Om te voorkomen dat studenten relevante informatie aan elkaar doorgeven moeten alle studenten op hetzelfde tijdstip geëxamineerd worden. Dit betekent dat meerdere simulatiepatiënten geprogrammeerd moeten zijn op hetzelfde ziektebeeld en dat meerdere observatoren moeten worden ingeschakeld om de verrichtingen te beoordelen. De standaardisatie wordt daardoor van twee kanten bedreigd:

- in de eerste plaats doordat het niet voor honderd procent zeker is dat elke simulatiepatiënt op dezelfde manier zal reageren op vragen, opmerkingen of onderzoek van de student;
- in de tweede plaats doordat beoordelaars dezelfde verschijnselen niet altijd hetzelfde beoordelen. In par. 2.2.2 van deel I

van dit proefschrift werden de problemen besproken van het inschakelen van beoordelaars.

Een ander nadeel van simulatiepatiënten, dat met name actueel is als studenten nog onervaren zijn, is dat zowel communicatieve als intellectuele vaardigheden aangewend dienen te worden tijdens het diagnostisch proces. Bij papieren simulaties, daarentegen, kan de student zich volledig concentreren op de cognitieve aspecten van het patiëntprobleem. Vanuit een oogpunt van training is dit soms aantrekkelijk.

Het grootste nadeel van het gebruik van simulatiepatiënten is dat ze geen oplossing bieden voor het probleem dat inherent is aan het gebruik van echte patiënten ten behoeve van onderwijsdoeleinden, namelijk dat studenten niet de kans krijgen om de consequenties te ervaren van beslissingen die zij in het oplosproces nemen. Papieren simulaties bieden deze mogelijkheid wel. In par. 1.5.2 wordt nader ingegaan op de voordelen van papieren simulatie en op enkele in het medisch onderwijs bekende vormen ervan. Paragraaf 1.5.3, tenslotte, zal geheel gewijd zijn aan de bespreking van één vorm van papieren simulatie: het patiënt management probleem.

1.5.2 Papieren simulatie

Papieren simulaties zijn beschrijvingen van situaties waarmee de beoefenaar van een bepaald beroep in de praktijk geconfronteerd kan worden. Wie de simulatie "doorwerkt" wordt geacht te denken en te beslissen alsof het een reële situatie betreft. Consequenties van beslissingen worden onmiddellijk kenbaar gemaakt aan degene die het gesimuleerde probleem probeert op te lossen. Wat papieren simulatie inhoudt laat zich het beste illustreren aan de hand van een bespreking van de vele voordelen die deze techniek biedt voor instructie- en evaluatie-doeleinden. McGuire (1976) noemt de volgende:

1. Omdat papieren simulaties de werkelijkheid beter benaderen dan examens die bestaan uit open of gesloten vragen, worden ze ook als relevanter gezien door de geëxamineerden. Deze waargenomen relevantie heeft een positieve invloed op de motivatie.
2. Papieren simulaties maken het mogelijk om een leer- of testsituatie te creëren die alleen dié elementen bevat die werkelijk van belang zijn. Met andere woorden: papieren simulaties imiteren de werkelijkheid maar zijn geen duplikaat daarvan, omdat de in de werkelijkheid altijd aanwezige "ruis" vermeden wordt. Door deze eigenschap wordt het mogelijk om studenten te confronteren met precies dezelfde taak en hun prestaties onderling te vergelijken.
3. Doordat de uit te voeren taak nauwkeurig gedefinieerd kan worden is het mogelijk om specifieke en gedetailleerde criteria te ontwikkelen waarmee de prestaties op de taak beoordeeld kunnen worden. Prestaties kunnen daardoor vrij objectief beoordeeld worden, hetgeen de waarde van de teruggekoppelde informatie ten goede komt.
4. Zelfs in uiterst kritieke situaties behouden studenten de volle

verantwoordelijkheid voor de gevolgen van hun beslissingen. Door studenten zelf te laten beslissen in kritieke situaties en hun vervolgens realistische terugkoppeling te geven over de gevolgen van hun beslissingen, is het mogelijk hen te trainen in "besluitvorming".

5. Papieren simulaties maken het mogelijk om ervaring op te doen met langdurige gebeurtenissen of met de lange termijn effecten van bepaalde acties in een oefening van ongeveer 30 minuten. Een medisch student kan zo ervaren wat de uitwerking is van zijn beslissing om, bijvoorbeeld, een gezwel operatief te verwijderen. De onmiddellijke, specifieke en ondubbelzinnige terugkoppeling die aan studenten gegeven kan worden maakt dat papieren simulaties uitstekende leermiddelen zijn.

In het medisch onderwijs was Rimoldi (1961) de eerste die probeerde om op objectieve wijze "diagnostische vaardigheid" te meten. Zijn "Test for Diagnostic Skills" bestaat uit een groot aantal kaartjes met vragen langs de bovenrand en antwoorden op de achterzijde. Op grond van schriftelijke informatie over een casus kiest de geëxamineerde een kaart waarop een vraag staat die hij zou willen stellen. Op de achterzijde van het kaartje wordt een antwoord op die vraag gegeven. De geëxamineerde neemt kennis van het antwoord en stelt op grond van de nieuw verworven informatie een nieuwe vraag. Dit gaat zó lang door totdat de geëxamineerde denkt voldoende informatie te hebben om een diagnose te kunnen stellen. De volgorde waarin de kaartjes gekozen zijn geven informatie over de wijze waarop het probleem benaderd is. Om een vergelijking van "probleemaanpak" tussen geëxamineerden eenvoudiger te maken, zijn alle kaartjes voorzien van een nummer.
Een vrij recente variant van Rimoldi's werkwijze staat bekend onder de naam "P4-deck" (Barrows en Tamblyn, 1980, geciteerd in Metz, 1984). P4 staat voor Portable Patient Problem Package. In dit pakket zijn de verschillende secties, bijvoorbeeld anamnese, onderzoek, behandeling, enz. gecodeerd met kleuren. Behalve het item langs de bovenrand van het kaartje, waarover op de achterzijde informatie wordt verstrekt, staan aan de voorkant een aantal vragen die de student zichzelf kan stellen. Bijvoorbeeld: "Wat verwacht je hiervan?"; "Belastend voor de patiënt?"; "Wat zijn de kosten?". Dergelijke vragen dwingen de student om gemotiveerd te kiezen voor een vraag, onderzoek, behandeling, enz.

Een in Nederland snel populair wordende methode voor het vaststellen van diagnostische vaardigheid is de Gestructureerde Open Vraag (GOV)(Metz, 1984). Naar aanleiding van een casusbeschrijving wordt een aantal korte open vragen geformuleerd. Elke keer nadat een vraag beantwoord is krijgt de respondent terugkoppeling over de juistheid van zijn antwoord en tevens informatie die weer leidt tot de volgende korte open vraag. Deze methode verenigt een belangrijk voordeel van "open vragen" met een belangrijk voordeel van "gesloten vragen". Zo zijn de antwoorden dermate kort dat objectief beoordelen beter mogelijk is dan bij gewone open vragen en is er geen sprake van "cueing".

Cueing doet zich voor als een probleemoplosser geconfronteerd wordt met een aantal alternatieven waaruit hij een keuze moet maken. De probleemoplosser kan de alternatieven tegen elkaar afwegen en een keuze maken die hij niet uit zichzelf overwogen zou hebben (Verdonschot, 1984).

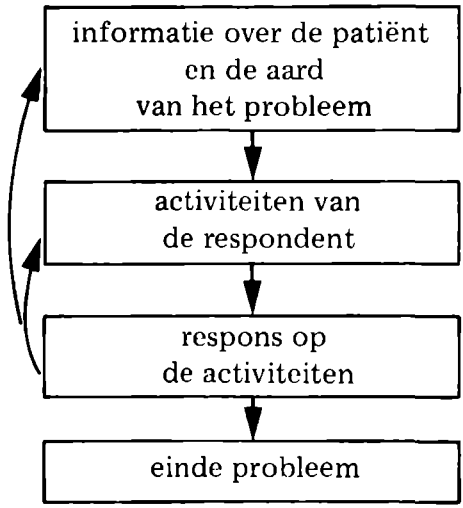
Het bekendste instrument waarmee praktijksituaties op papier gesimuleerd kunnen worden is het Patiënt Management Probleem (PMP)(De Graaff en Galesloot, 1982). Een PMP vangt aan met een beknopte, maar realistische casusbeschrijving. Uit de beschrijving wordt ook duidelijk onder welke omstandigheden het probleem opgelost moet worden, welke rol de probleemoplosser moet aannemen en wat zijn taak is (McGuire, 1976). Het probleem moet worden opgelost door zó veel informatie in te winnen (door vragen te stellen aan de patiënt, door het verrichten van lichamelijk onderzoek of laboratoriumonderzoek) dat een diagnose gesteld kan worden en een eventuele behandeling kan worden uitgevoerd. Deze informatie is opgeslagen in items (opties met bijbehorende responsen), die gegroepeerd zijn in secties. Als voor een bepaalde optie gekozen is krijgt de oplosser de beschikking over de respons. Het verstrekken van deze terugkoppeling kan op verschillende manieren georganiseerd worden. In par. 1.5.3.1 worden de meest voorkomende besproken.

Voor het vaststellen van probleemoplosvaardigheid is het PMP geschikter dan de besproken methode van Rimoldi, het P4-deck en de gestructureerde open vraag. De "kaartjes-methode" van Rimoldi en het P4-deck zijn met name minder geschikt als grote aantallen studenten gelijktijdig getoetst moeten worden. Om te voorkomen dat een student nadat deze de respons op de achterzijde van een kaartje gelezen heeft, dit kaartje weer terugstopt in geval van een verkeerde keuze, is in een toetssituatie intensieve controle op fraude noodzakelijk. De gestructureerde open vraag heeft als belangrijkste nadeel dat ze niet erg realistisch is. Oplossers worden niet gedwongen om eenmaal gemaakte keuzes uit te werken en de consequenties van genomen beslissingen te dragen. Na elk antwoord wordt terugkoppeling gegeven in de vorm van het juiste antwoord (volgens experts). De methode staat niet toe dat van de uitgestippelde oplosroute wordt afgeweken.
Voor het vaststellen van (medische) probleemoplosvaardigheid van grote aantallen studenten lijkt het PMP daarom het meest geschikte instrument. In de volgende paragraaf wordt uitvoeriger aandacht besteed aan het PMP.

1.5.3 Patiënt Management Problemen

1.5.3.1 De structuur van PMP's

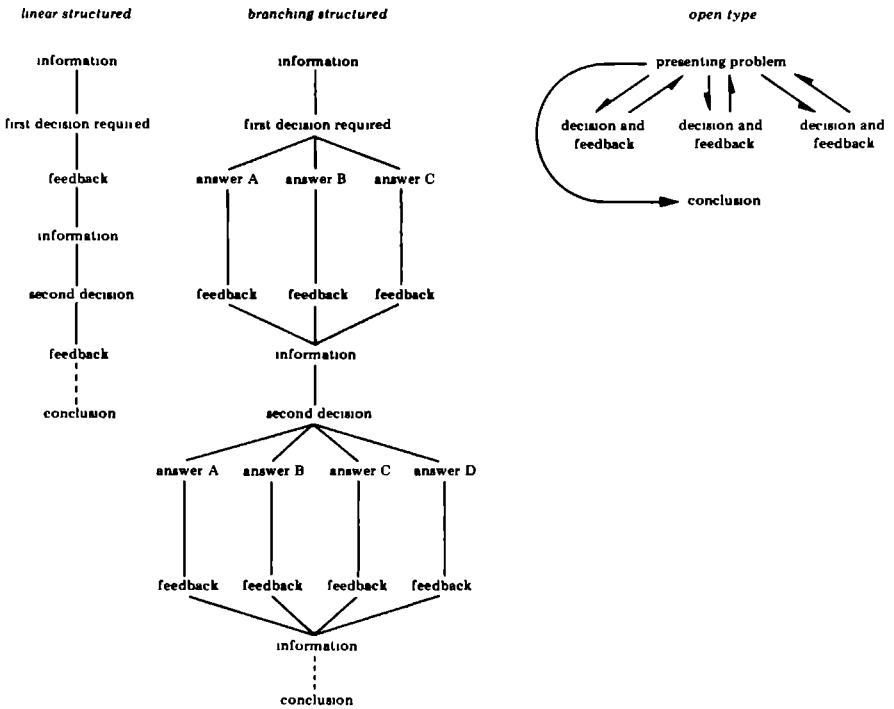
Sinds McGuire in 1967 voor het eerst experimenteerde met PMP's zijn er vele variaties gekomen op hetzelfde idee. Aan al deze variaties ligt dezelfde structuur ten grondslag, die schematisch is weergegeven in figuur 1.1.



Figuur 1.1: Fasen in een PMP (Harden, 1983).

De eerste fase, waarin informatie wordt verstrekt over de patiënt en zijn probleem, is voor alle varianten ongeveer op gelijke wijze ingericht.
Dit geldt niet voor de tweede fase waarvoor in essentie drie varianten bestaan: de lineaire, de vertakte en de open variant (zie figuur 1.2).
Bij een lineair probleem wordt de respondent gedwongen om een vastgesteld aantal beslissingen in een voorgeschreven volgorde te nemen. De oplosroute zal dus voor elke respondent gelijk zijn, waarmee het beperkte nut is aangegeven van dergelijke PMP's voor het vergelijken van de manier waarop problemen worden aangepakt door respondenten.
Evenals het lineaire probleem confronteert het vertakte probleem alle respondenten met dezelfde te nemen beslissingen. Maar de route naar elke beslissing is afhankelijk van de eerder genomen beslissing. McGuire (1976) noemt deze variant de "forced-branching technique".
In het open probleem bepaalt de respondent zelf op welke wijze hij door het probleem gaat. Dit wordt bewerkstelligd door in elke sectie een "brugsegment" op te nemen bestaande uit verschillende "takken" en een instructie voor respondenten om de "tak" van hun voorkeur te kiezen. McGuire (1976) onderscheidt twee vormen van het open probleem:
- de vrij vertakte PMP's;
- de gemodificeerd vrij vertakte PMP's.
De vrij vertakte PMP's stellen respondenten in staat om geheel vrij van sectie naar sectie te gaan in een door henzelf gewenste volgorde. Gemodificeerd vrij vertakte PMP's laten de respondent in principe net zo vrij als de normale vrij vertakte PMP's, maar hebben een voorziening waarmee vroegtijdige beëindiging wordt

voorkomen. De vertakkingen in brugsegmenten die, als ze gekozen worden, in een vroegtijdige beëindiging zouden resulteren, worden geblokkeerd door de respondent te dwingen om een andere keuze te maken.



Figuur 1.2: Drie soorten van Patiënt Management Problemen. (Harden, 1983)

Tot zover is alleen gesproken over de volgorde van activiteiten die respondenten ondernemen om het voorgelegde probleem op te lossen. Naar de aard van de te ondernemen activiteiten kan ook onderscheid gemaakt worden. Harden (1983) noemt de volgende:

1. het inwinnen van meer informatie door:
   - het opnemen van een anamnese;
   - het verrichten van lichamelijk onderzoek;
   - het verrichten van laboratorium- en/of radiologisch onderzoek.
2. het interpreteren van de verworven informatie door:
   - een lijst op te stellen van de gevonden problemen;
   - conclusies te trekken op grond van bepaalde bevindingen;
   - een hypothese of diagnose te stellen.
3. het behandelen van de patiënt door:
   - advies te geven;

- medicijnen voor te schrijven;
- dieet voor te schrijven;
- te verwijzen naar een specialist;
- regelmatige controle voor te schrijven.

In de derde fase van het PMP wordt terugkoppeling verstrekt aan de respondent over een of andere ondernomen actie. De terugkoppeling kan van tweeërlei aard zijn:

1. terugkoppeling in de zin van het geven van informatie over het al dan niet juist zijn van de ondernomen activiteit;
2. terugkoppeling in de zin van het verstrekken van dié informatie die in een werkelijke situatie ook verkregen zou zijn bij het uitvoeren van bepaalde activiteiten.

Meestal zal er in een PMP gebruik worden gemaakt van beide vormen van terugkoppeling.

PMP's kunnen geheel zelfstandig opgelost worden door de respondenten voor wie ze geconstrueerd zijn. Dit betekent dat de terugkoppeling in het PMP zelf opgenomen moet zijn. Maar om voortijdige verstrekking van terugkoppeling te voorkomen moet deze op een of andere manier "verborgen" worden. Harden (1983) geeft een opsomming van mogelijkheden waarop dit kan gebeuren.

1. De terugkoppeling is afgedrukt op de volgende pagina. De respondent mag niet eerder naar een nieuwe pagina voordat schriftelijk een antwoord is gegeven op de gestelde vraag.
2. De terugkoppeling is op dezelfde pagina afgedrukt, gescheiden door een lijn van de vraag of in een tweede kolom naast de betreffende vraag.
3. De terugkoppeling is afgedrukt op een willekeurige pagina waarnaar verwezen wordt door een nummer achter de vraag.
4. De terugkoppeling is onzichtbaar afgedrukt in een tweede kolom achter de betreffende vraag (latent image printing). Met een speciale pen kan de informatie zichtbaar gemaakt worden.
5. De terugkoppeling is onleesbaar gemaakt doordat er strepen in een andere kleur over zijn afgedrukt. Een transparant in dezelfde kleur als de strepen maakt de informatie weer leesbaar.
6. De terugkoppeling is verborgen onder een ondoorzichtige bedekking die door schrappen verwijderd kan worden.

De onzichtbare druktechniek is in vele opzichten te verkiezen boven andere technieken. De volgende argumenten kunnen hiervoor aangevoerd worden:

- respondenten komen niet in de verleiding om "stiekem" te kijken;
- eenmaal genomen beslissingen kunnen niet meer ongedaan gemaakt worden;
- respondenten hoeven niet zelf antwoorden te formuleren; waardoor objectieve beoordeling eenvoudiger realiseerbaar is;
- het is mogelijk om oplosroutes te reconstrueren;

Aan de techniek zijn ook nadelen verbonden:

- de drukkosten zijn hoger dan bij "normaal" drukwerk;
- de PMP's kunnen slechts één keer gebruikt worden;
- de kwaliteit van de onzichtbaar gedrukte teksten is onzeker tot

op het moment dat ze ontwikkeld worden.
Het nadeel van de eenmalige bruikbaarheid is vermeden bij een nog in ontwikkeling zijnde techniek, waarbij cholesterische kristallen gebruikt worden. Deze reageren op temperatuur en onthullen informatie tijdelijk als de temperatuur stijgt als gevolg van het wrijven met een duim erover (Cairncross en Harden, 1983, geciteerd in Harden, 1983). Een nadeel van deze techniek is dat ze de besluitvorming van de respondent niet vastleggen. Deze techniek is daarom alleen geschikt als de PMP's voor instructie-doeleinden of zelf-evaluatie gebruikt worden.

De vierde en laatste fase in een PMP betreft de conclusie. In een gestructureerd probleem komt dit overeen met de laatste vraag. In een open probleem met het voltooien van de opgelegde taak. Als het een diagnostisch probleem betreft is de taak volbracht als de respondent denkt dat hij voldoende informatie heeft verzameld om een diagnose te kunnen stellen. In een behandelingsprobleem als de patiënt herstelt of overlijdt.

1.5.3.2 Enkele nadelen van PMP's

De voordelen die het gebruik van PMP's oplevert zijn al uitvoerig besproken in par. 1.5.2 waar een opsomming werd gegeven van de voordelen van papieren simulatie in het algemeen. In dezelfde paragraaf werd geconcludeerd, dat PMP's de voorkeur genieten boven andere vormen van papieren simulatie vanwege het feit dat ze beter geschikt zouden zijn om grote aantallen studenten gelijktijdig te toetsen en vanwege hun grote natuurgetrouwheid. Een bespreking van de aan PMP's verbonden nadelen mag hier echter niet achterwege blijven. De belangrijkste zijn:
1. de constructie neemt veel tijd in beslag;
2. het blijkt bijzonder moeilijk om een bevredigend scoringssysteem te ontwikkelen;
3. er wordt getwijfeld aan de validiteit van PMP's.

ad.1 Moeizame constructie

Het construeren van een PMP is een moeizame en zeer tijdrovende bezigheid. Met name als het gaat om een "open probleem" heeft een constructeur veel werk om alle keuzemogelijkheden uit te werken in een voor de respondent realistisch perspectief. McGuire (1976) schat dat een constructeur gemiddeld 40 tot 50 uur nodig heeft om zijn eerste, 30 tot 40 minuten durende, simulatie te schrijven. Ervaren constructeurs zouden volgens haar slechts 5 tot 10 uur nodig hebben voor een simulatie van die omvang. De constructietijd kan teruggebracht worden door het opstellen van constructieschema's voor de terzake-kundigen die de PMP's inhoudelijk vorm moeten geven. Verdonschot (1983) ontwikkelde een geprogrammeerde instructie voor het construeren van tandheelkundige PMP's. In Leiden is onderzoek verricht naar de mogelijkheid om met behulp van een computerprogramma vooraf gespecificeerde patiëntgegevens te verwerken tot een computerpatiënt (Verbeek, 1982).

Hoewel dergelijke ontwikkelingen van nut zijn om de constructiewerkzaamheden te vereenvoudigen, moet het leeuwedeel van de werkzaamheden toch nog door de constructeur zelf verricht worden. Van de constructeur worden creativiteit en vakkennis verwacht om PMP's voldoende realistisch te kunnen laten zijn om gemotiveerde inzet van respondenten te garanderen.

ad.2 Ingewikkelde scoring

Met name als PMP's gebruikt worden als middel om beslissingen te nemen over studenten, is een scoringssysteem onontbeerlijk. Bij papieren simulaties moet de scoring anders van aard zijn dan gebruikelijk is bij de meeste schriftelijke toetsen. Bij laatstgenoemde evaluatie-instrumenten wordt in de meeste gevallen alleen met positieve itemscores gewerkt. Bij een goed antwoord wordt een punt of worden meerdere punten toegekend, bij een foutief antwoord of het onbeantwoord laten gebeurt er niets. Wil men in een toets- of oefensituatie de werkelijkheid simuleren, dan moet dit onder andere ook tot uiting komen in het straffen van verkeerde beslissingen. Vandaar dat negatieve scoring bij PMP's heel gewoon is. Om beter te kunnen discrimineren tussen "goede", "minder goede" en "slechte" probleemoplossers wordt vaak gebruik gemaakt van gewogen scoring. Afhankelijk van het belang van een beslissing voor de oplossing van het probleem variëren de scores van bijvoorbeeld -3 (voor beslist te vermijden keuzes) tot +3 (voor beslist essentiële keuzes). Maar aan dit ogenschijnlijk gedegen scoresysteem kleven de volgende nadelen:

1. Marshall (1977) ontdekte dat de meest efficiënte oplossers niet altijd de hoogste scores behaalden. Oplossers die de goede oplossing via een omweg bereikten bleken hogere scores te kunnen behalen dan de "kortste weg" oplossers, doordat ze meer opties kozen en dus meer bonuspunten konden verzamelen. Hij stelde voor dit probleem op te lossen door voor elke sectie een maximum score vast te stellen, gebaseerd op het scoretotaal van de in de betreffende secties belangrijke items. Dit systeem bleek bevredigend te werken.
2. De validiteit van de scores is afhankelijk van de validiteit en betrouwbaarheid van de weging. De validiteit van de weging is een vakinhoudelijke kwestie. Of de weging betrouwbaar is, is afhankelijk van de opinies van de deskundigen over het belang van het te wegen item voor de oplossing van het probleem (Metz, 1984). Door verschillen in opvatting, opleiding, kennis, enz. zal het bijzonder moeilijk zijn om hoge overeenstemming te bereiken tussen de deskundigen over de toe te kennen gewichten. Een mogelijke oplossing ter vermijding van deze validiteitsproblemen ligt in het gebruik van descriptieve (in tegenstelling tot evaluatieve) scoringssystemen. Deze scoringssystemen zijn niet bedoeld om de oplosvaardigheid in één of meerdere scores uit te drukken, maar om met behulp van bepaalde indices aan te geven hoe het probleem in kwestie is aangepakt (Metz, 1984). Evaluatieve scoringssystemen zijn vooral van belang voor het nemen van beslissingen over studenten, terwijl descriptieve scoringssystemen vooral van nut zijn bij het nemen

van beslissingen over het onderwijs en het evaluatie-instrument zelf. Dikwijls zal er behoefte bestaan aan beide vormen van scoring, zodat de tijdrovende wegingsprocedures van de evaluatieve scoring toch noodzakelijk zijn.

ad.3 Validiteitsprobleem

PMP's simuleren het besluitvormingsproces dat zich afspeelt als de arts een diagnose stelt om op grond daarvan het geïdentificeerde probleem op te lossen. Een belangrijke vraag is of het behaalde resultaat op een PMP representatief kan zijn voor de klinische situatie. Met andere woorden: kan klinische probleemoplosvaardigheid gemeten worden met behulp van PMP's? Om deze vraag te beantwoorden vergeleken Goran et al. (1973) het klinisch oordeel van artsen met het oordeel van dezelfde artsen over een analoog probleem in de context van een PMP. Het gedrag op het PMP week significant af van het gedrag in de klinische situatie. In het PMP werden:
- meer relevante gegevens opgevraagd;
- diagnoses zorgvuldiger gesteld;
- meer laboratoriumonderzoeken aangevraagd.

De auteurs verklaarden deze resultaten als volgt:
- In PMP's spelen tijd en economische factoren een minder belangrijke rol dan in een klinische situatie. Het is erg gemakkelijk om in de context van een PMP informatie op te vragen en onderzoek te doen.
- Bij PMP's is het niet mogelijk om met zekerheid vast te stellen of de respondent werkelijk over de kennis beschikt om relevante vragen te kunnen stellen. De respondent hoeft alleen maar uit een lijst de meest geschikte vragen te selecteren. Dit effect staat bekend onder de naam "cueing" (zie par. 1.5.2).
- In een klinische situatie moet een arts eerst uitzoeken welke informatie zinvol zou kunnen zijn en vervolgens deze informatie zien te bemachtigen. In een PMP hoeft een respondent alleen maar vast te stellen welke informatie belangrijk is om de beschikking daarover te krijgen. Het PMP meet alleen maar de intentie om onderzoek te doen en niet de bekwaamheid om de antwoorden van de patiënt juist te kunnen interpreteren of om lichamelijk onderzoek uit te voeren.

Op grond van de resultaten concluderen de auteurs dat getwijfeld moet worden aan de validiteit van PMP's om klinisch denken te meten.

Page en Fielding (1980) verrichtten eveneens onderzoek naar de criteriumvaliditeit van PMP's. Zij vergeleken de prestaties van apothekers op PMP's met hun prestaties op een analoog probleem in de praktijk. Voor dit doel werd gewerkt met uitvoerig getrainde acteurs. Voorafgaand aan het onderzoek werden de PMP's getest op inhoudsvaliditeit en constructvaliditeit. Inhoudsvaliditeit werd vastgesteld toen tien experts van mening waren dat de PMP's prima representaties waren van de in de praktijk-situatie aangeboden casussen. Om constructvaliditeit vast te kunnen stellen werden bij eerstejaars psychologie studenten, eerstejaars pharmacie

studenten, ouderejaars pharmacie studenten en apothekers alle vier geconstrueerde PMP's afgenomen. De prestaties van de vier groepen werden met elkaar vergeleken. Op één na waren alle verschillen in de verwachte richting: de psychologie studenten scoorden steeds het laagst, de apothekers het hoogst en de ouderejaars pharmacie-studenten hoger dan hun collega's uit het eerste jaar. Vergelijking tussen prestaties op PMP's en prestaties in praktijk-situaties wees uit dat in PMP's gemiddeld achttien procent meer activiteiten ontplooid werden dan in de vergelijkbare praktijk-situaties. De auteurs denken dat dit verschil voornamelijk veroorzaakt wordt door het cueing-effect en door het in conflict komen met elkaar van de professionele en zakelijke rol van de apotheker. Dit laatste zou volgens de auteurs te maken hebben met het feit dat PMP's "prestatiebekwaamheid" meten en niet de prestatie zelf. Meer gedetailleerde analyses resulteerden in de constatering dat PMP's goede voorspellers waren van wat apothekers in praktijk-situaties niet zouden doen en slechte voorspellers voor wat ze wel zouden doen. De voorspelling was met name slecht voor gedragingen van het type "beslist te kiezen" en "beslist te vermijden". Op grond van deze resultaten twijfelen de auteurs aan de criteriumvaliditeit van de bestudeerde PMP's.

Norman en Feightner (1981) vergeleken de prestaties van studenten op simulatie-patiënten met hun prestaties op inhoudelijk vergelijkbare PMP's. Elke student loste twee PMP's op en nam een interview af bij twee simulatie-patiënten. De belangrijkste conclusie was dat het oplossen van PMP's gepaard ging met het opvragen van meer informatie. De auteurs vragen zich af of PMP's wel gebruikt kunnen worden als een maat voor medische competentie.

Terecht waarschuwt Marshall (1983) ervoor om PMP's op grond van de resultaten uit de hier besproken onderzoeken te beschouwen als nutteloze instrumenten. Hij wijst op de fout die de onderzoekers gemaakt hebben door te veronderstellen dat prestaties op PMP's, simulatie-patiënten en echte patiënten een maat zijn voor identieke gedragingen. Maar het PMP meet slechts "probleemoplosvaardigheid" en geen "attitudes" of "psychomotorische vaardigheden". Criteriumvaliditeit kan daarom niet goed vastgesteld worden door prestaties op PMP's te vergelijken met prestaties op simulatie- of op echte patiënten. Als het gaat om het vaststellen van de bekwaamheid om problemen van medische aard op te lossen, zijn PMP's, door hun hoge inhoudsvaliditeit en door hun vermogen (mits gebruik wordt gemaakt van een goed scoringssysteem) om onderscheid te maken tussen goede en minder goede probleemoplossers, als zeer geschikt te beschouwen.

1.6 Discussie

Het voornaamste kenmerk van probleemgeoriënteerd onderwijs is dat de nadruk op afzonderlijke vakken plaats heeft gemaakt voor een meer geïntegreerde benadering, waardoor studenten (naar verwachting) beter worden voorbereid op de latere beroepsuitoefe-

ning. Dit wordt bewerkstelligd door studenten te laten werken aan problemen. De nadruk ligt daarbij meer op de wijze waarop het probleem wordt aangepakt dan op de oplossing zelf. Probleemoplosvaardigheid is meer dan alleen het kunnen geven van de oplossing. In zijn algemeenheid kan als eis gesteld worden dat het oplosproces efficiënt dient te verlopen. Dit betekent dat, indien mogelijk, gebruik moet worden gemaakt van algorithmen en dat in alle andere gevallen geprobeerd moet worden om met gebruik van de meest geschikte heuristische procedures de oplossing op systematische wijze te bereiken.
Het vaststellen van probleemoplosvaardigheid bij grote aantallen studenten is om de volgende redenen problematisch:
- methodes als "het maken van hardop-denkprotocollen" en "het observeren van het oplosproces" kunnen door hun arbeidsintensieve karakter niet efficiënt worden toegepast;
- in vakgebieden waar de problemen van menselijke aard zijn is het praktisch onmogelijk om gestandaardiseerde problemen aan te bieden. Voor het onderwijs heeft dit tot gevolg dat moeilijk vastgesteld kan worden wie de doelstellingen bereikt heeft en wie (nog) niet.

Het gebruik van papieren simulatie als methode om probleemoplosvaardigheid vast te stellen biedt in genoemde omstandigheden een uitkomst. Met name in het medisch onderwijs is veel ervaring opgedaan met deze methode. Drie bekende vormen van papieren simulatie in het medisch onderwijs zijn de "kaartjes-methode", de gestructureerde open vraag en het patiënt management probleem.
De kaartjes-methode stelt de oplosser in staat om door het kiezen van kaartjes vragen te stellen aan de patiënt, onderzoek te doen of de patiënt te behandelen. De volgorde waarin de kaartjes gekozen worden geeft informatie over de gevolgde probleemaanpak. Omdat het eenvoudig is om heimelijk terugkoppeling te krijgen is deze methode niet geschikt om bij grote aantallen studenten als formele toets gebruikt te worden.
De gestructureerde open vraag is een methode waarbij naar aanleiding van een casusbeschrijving een reeks korte open vragen beantwoord moet worden. Als een student een vraag beantwoord heeft krijgt hij onmiddellijk terugkoppeling en informatie over hoe de casus zich verder ontwikkelt. Het probleem met deze methode is de subjectieve beoordeling van de antwoorden en het niet zo realistische karakter.
Het patiënt management probleem vangt eveneens aan met een korte casusbeschrijving en wordt opgelost door telkens uit een aantal gegeven items een keuze te maken, waardoor informatie verkregen wordt die de oplossing al dan niet dichterbij brengt. De nadelen van de twee hiervoor besproken methoden vormen juist de sterke punten van het patiënt management probleem. PMP's zijn zeer geschikt om grote aantallen studenten gelijktijdig te toetsen, ze zijn erg realistisch en objectief scoorbaar.

Behalve als toetsmiddel kunnen papieren simulaties dienst doen als leermiddel. Door middel van papieren simulaties is het mogelijk om al in een zeer vroeg stadium van de opleiding probleemgeoriënteerd te werken. Zonder dat personen gevaar lopen kunnen studenten

zelfstandig werken aan de oplossing van problemen en de gevolgen van hun beslissingen leren overzien.
Doordat gestandaardiseerde problemen aangeboden kunnen worden is het voor docenten mogelijk om na te gaan welke onderdelen van de leerstof moeilijkheden opleveren voor de meeste studenten. Hieraan kan dan in een klassikale instructie aandacht worden besteed. Het grote voordeel van deze werkwijze is dat studenten al in de beginfase van hun opleiding kunnen ervaren wat het belang is van de diverse vakken voor het oplossen van reële problemen.

De toepassing van papieren simulatie beperkt zich niet tot het medisch onderwijs. In hoofdstuk II wordt besproken hoe PMP's zinvol kunnen worden ingeschakeld in een preklinische cursus "behandelingsplanning" in het tandheelkundig onderwijs. De voor die cursus geconstrueerde PMP's zijn gebaseerd op een recent ontwikkeld probleemoplossingsmodel voor tandheelkundige problemen (Verdonschot, 1984). In dit model spelen heuristische procedures als "middel-doel analyse" en "fractionering" een belangrijke rol. De vervaardigde PMP's kunnen behalve als leer- ook als toetsmiddel gebruikt worden. In dat laatste geval dienen ze ter vervanging van de zogenaamde "papieren patiënt problemen", waarmee tot op heden probleemoplosvaardigheid wordt vastgesteld en waaraan verschillende ernstige bezwaren kleven.

II HET VASTSTELLEN VAN PROBLEEMOPLOSVAARDIGHEID IN HET PREKLINISCH TANDHEELKUNDIG ONDERWIJS

2.1 Inleiding

Met de curriculumherziening van de Nijmeegse Subfaculteit Tandheelkunde in 1974 heeft thematisch gericht onderwijs bijzondere aandacht gekregen. Het onderwijs wordt gegeven in blokcursussen waarin steeds één thema centraal staat. Er wordt naar gestreefd om in elke blokcursus "probleemgeoriënteerd" te werken, wat inhoudt dat de nadruk op het verwerven en reproduceren van kennis minder is geworden ten gunste van de toepasbaarheid van die kennis. Geconfronteerd met een concreet tandheelkundig probleem leren studenten hoe ze reeds verworven cognitieve, psychomotorische en affectieve vaardigheden geïntegreerd moeten toepassen om tot een oplossing te komen.
De tandheelkundige problematiek is bij patiënten vaak dermate gecompliceerd, dat de tandarts een draaiboek moet maken voor de behandeling. Zo'n draaiboek wordt een "behandelingsplan" genoemd en wordt door Käyser (1981) omschreven als "...een geordend rapport waarin alle van belang zijnde informatie met betrekking tot de klachten en afwijkingen van de patiënt naar bevinding, oorzaak, ernst, uit te voeren behandeling en te verwachten resultaat, wordt vermeld. Het doel hiervan is om op inzichtelijke wijze aan te geven waarom een bepaalde behandeling wordt geadviseerd, op welke wijze die moet worden uitgevoerd en wat de toekomstverwachting is. Tevens geeft het plan inzicht in de benodigde tijdsduur, de eventuele laboratoriumfasen en de kosten". Studenten worden expliciet onderwezen in het op systematische wijze opstellen van behandelingsplannen.

In dit hoofdstuk zal achtereenvolgens aandacht besteed worden aan:
- een nieuw ontwikkelde methode voor het systematisch opstellen van behandelingsplannen (par. 2.2);
- de nadelen van de vigerende methode voor het vaststellen van de bekwaamheid om behandelingsplannen op te stellen (par. 2.3);
- de ontwikkeling van een nieuwe methode voor het vaststellen van de bekwaamheid om behandelingsplannen op te stellen (par. 2.4).

2.2 Een nieuwe methode voor het opstellen van behandelingsplannen

Het opstellen van behandelingsplannen wordt in de Nijmeegse Subfaculteit Tandheelkunde onderwezen in het onderwijsblok "klinische tandheelkunde conserverend I" (blok 261). Aan studenten wordt geleerd om tandheelkundige problemen van patiënten te benaderen met een modificatie van wat door De Groot (1971) is aangeduid met de term "empirische cyclus". Volgens De Groot kunnen in het empirisch denken en onderzoeken de volgende fasen onderscheiden worden:

- observatie: het systematisch waarnemen, verzamelen en groeperen van empirisch feitenmateriaal;
- inductie: het opstellen van hypothesen;
- deductie: uit de theorie worden speciale, meer specifieke (en daardoor beter toetsbare) hypothesen afgeleid;
- confirmatie: de afgeleide hypothesen worden getoetst;
- evaluatie: de onderzoeker bezint zich op de betekenis en het nut van zijn onderzoek en de resultaten daarvan en beslist of hij het onderzoek als afgerond beschouwt of verder zal gaan.

De aanpassing van de empirische cyclus aan het specifieke domein van de tandheelkunde resulteerde uiteindelijk in de "gemodificeerde probleemoplossingscyclus", zoals die vanaf het studiejaar 1978-1979 in gebruik is. De volgende fasen worden daarin onderscheiden (Instituut Conserverende Tandheelkunde voor Volwassenen, 1978):

1. waarneming: alle belangrijke gegevens worden verzameld;
2. herkenning: tussen de waargenomen verschijnselen worden verbanden gelegd waarmee het ontstaan van de afwijking en de oorzaak ervan misschien verklaard kunnen worden;
3. probleemstelling: a. inventarisatie van alle aanwezige problemen;
   b. inperking tot relevante problemen;
4. mogelijke en te kiezen oplossingen: voor alle relevante problemen worden oplossingen opgesomd (liefst meerdere per probleem). Uit de oplossingen worden keuzes gemaakt die het beste bij de patiënt passen. Als standaardoplossingen niet mogelijk zijn worden modificaties gepland. In deze fase worden de geplande verrichtingen ook in volgorde van behandeling geplaatst en schattingen gemaakt van de benodigde tijd en de kosten.
5. toetsing concept behandelingsopzet: de gekozen oplossingen worden getoetst aan de meningen van specialisten en aan de opvattingen van de patiënt zelf;
6. definitieve behandelingsopzet: alle geplande verrichtingen staan in de juiste volgorde en worden gecompleteerd door ramingen van de benodigde tijd (aantal zittingen) en de kosten;
7. therapie: de verrichtingen worden volgens planning uitgevoerd;
8. prognose en evaluatie: de verwachting over de behandeling naar functie en duurzaamheid wordt vermeld. Er worden criteria opgesteld aan de hand waarvan de therapie beoordeeld wordt. Op grond van deze beoordeling kunnen nieuwe problemen herkend worden, waarmee de probleemoplossingscyclus opnieuw gestart wordt.

Het belangrijkste nadeel van de gemodificeerde probleemoplossingscyclus is dat hij studenten geen denkregels verschaft, volgens welke zij van de ene fase naar de andere kunnen komen. De probleemoplossingscyclus is daarom te karakteriseren als een aantal vuistregels waarmee een patiëntprobleem "aangepakt" kan worden, maar niet met zekerheid wordt opgelost. Verdonschot (1982) onderzocht of het mogelijk was om een oplosmethode te construeren

waarmee tandheelkunde studenten in staat zouden zijn om op systematische wijze geïntegreerde tandheelkundige problemen op te lossen. Bij de constructie van zijn probleemoplossingsmodel stonden de volgende uitgangspunten centraal:

1. Een patiënt is een complex probleem. Dit betekent dat voor de oplossing geen standaardmethode voorhanden is. Een probleemoplosmethode moet aan oplossers regels verstrekken waarmee het complexe probleem opgedeeld kan worden in meerdere, eenvoudiger op te lossen deelproblemen. Deze heuristische procedure is in par. 1.4 besproken onder de naam "fractionering".
2. Voor de oplossing van elk deelprobleem wordt een heuristische methode aangewend die in par. 1.4 is aangeduid met de term "middel-doel analyse". Om de oplosser te helpen het doel te formuleren en een geschikt "middel" te vinden om het te bereiken, worden denkregels verstrekt. Denkregels geven de probleemoplosser instructies met betrekking tot de wijze waarop ze van de ene tussenfase naar de andere kunnen komen.

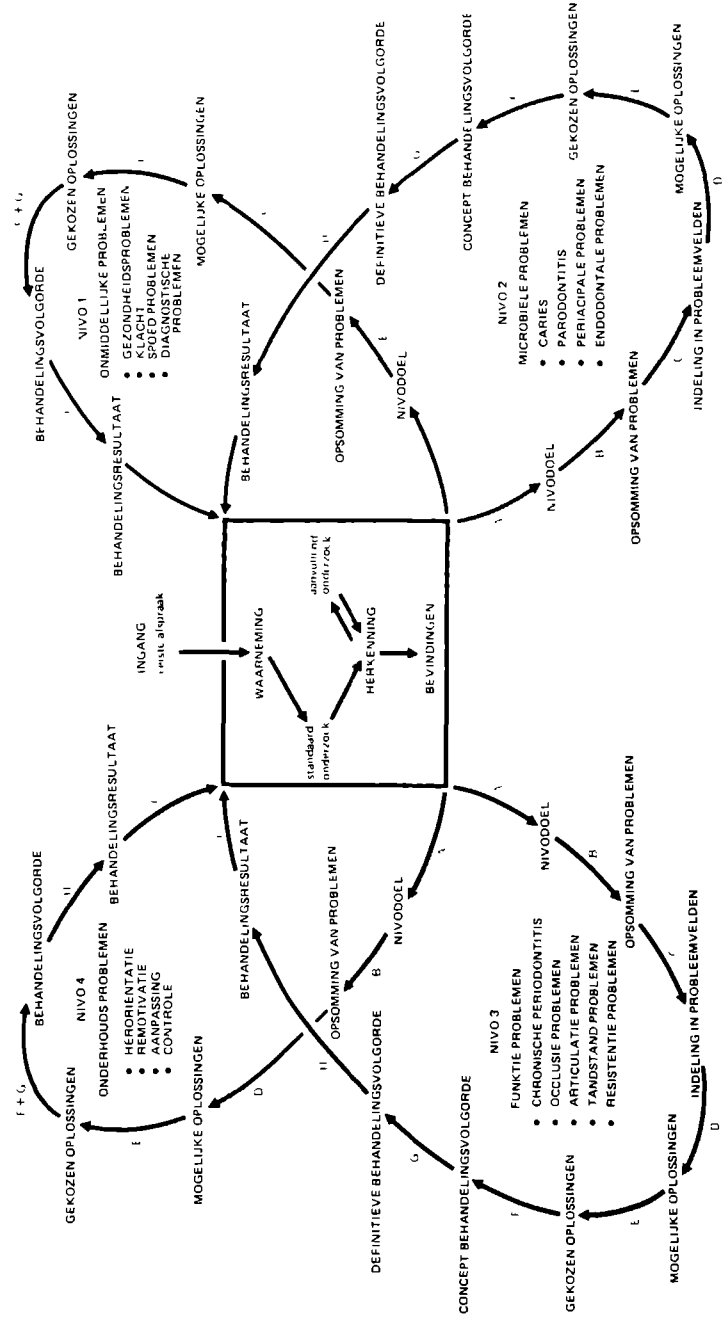
Het criterium voor de fractionering van het complexe patiëntprobleem is de urgentie waarmee aanwezige problemen moeten worden opgelost. Het probleemoplossingsmodel heeft vier niveau's die problemen bevatten van verschillende urgentie-graad. Het eerste niveau bevat problemen waaraan de hoogste prioriteit verleend moet worden. Het betreft de zogenaamde "onmiddellijke problemen", waartoe gerekend moeten worden: gezondheidsproblemen, de directe klacht, spoedproblemen en diagnostische onzekerheid. Problemen die direct daarna moeten worden opgelost zijn in het probleemoplossingsmodel geordend onder de noemer "microbiële problemen" (niveau 2). Het bekendste microbiële probleem is cariës, maar ook ontstekingen aan het parodontium en endodontische problemen worden hier toe gerekend. Als de microbiële problemen zijn opgelost kan de aandacht gericht worden op het herstel van verloren gegane functies (niveau 3: "functie-problemen"). Zaken als occlusie, articulatie, gebitsregulatie en gebitsresistentie komen daarbij aan de orde. Als laatste in de rij worden de "onderhoudsproblemen" (niveau 4) opgelost. Van dergelijke problemen is bijvoorbeeld sprake als de motivatie van de patiënt om zijn gebit te onderhouden niet groot genoeg is, of als door wijzigingen in de gebitssituatie de poetsmethode aangepast moet worden.

Binnen elk niveau moet de oplosser op zo efficiënt mogelijke wijze de waargenomen problemen oplossen en het resultaat daarvan evalueren. Voor dit doel is het oplosproces verdeeld in een aantal tussenfasen, waarin de oplosser activiteiten moet ontplooien die hem steeds dichter bij de oplossing brengen. Om de overgang van de ene fase naar de andere te vergemakkelijken worden denkregels gebruikt. In tabel 2.1 wordt de koppeling tussen denkregels en tussenfasen geïllustreerd.

Tabel 2.1: Koppeling tussen denkregels en tussenfasen in het probleemoplossingsmodel (Verdonschot, 1982).

DENKREGEL	resulteert in ⟹	TUSSENFASE
1. Vat de waargenomen problemen samen.		OPSOMMING VAN DE PROBLEMEN
2. Vertaal de problemen in termen van eindresultaten.		NIVEAUDOEL
3. Zoek gemeenschappelijke kenmerken bij de problemen en rangschik deze naar aard.		INDELING IN PROBLEEMVELDEN
4. Genereer voor elk probleem een of meer oplossingen.		MOGELIJKE OPLOSSINGEN
5. Kies uit de mogelijke oplossingen de meest juiste oplossingen.		GEKOZEN OPLOSSINGEN
6. Beredeneer de keuze en rangschik de problemen naar de volgorde waarin deze behandeld moeten worden.		CONCEPT BEHANDELINGSVOLGORDE
7. Beoordeel de concept-behandelingsvolgorde kritisch en overleg met de patiënt.		DEFINITIEVE BEHANDELINGSVOLGORDE
8. Voer de geplande verrichtingen uit conform de definitieve behandelingsvolgorde.		BEHANDELINGSRESULTAAT
9. Evalueer het behandelingsresultaat door het af te wegen tegen het gestelde niveaudoel.		WAARNEMING

In het probleemoplossingsmodel neemt de "waarneming" een centrale plaats in. Van hieruit starten en eindigen de oplosprocessen in de verschillende niveaus. Duidelijk zichtbaar is dit in figuur 2.1, dat een schematische weergave is van het probleemoplossingsmodel. De vraag of tandheelkundige problemen beter worden opgelost als gebruik gemaakt wordt van het probleemoplossingsmodel, is tot op heden onbeantwoord gebleven. Slechts op zeer beperkte schaal zijn onderdelen van het model getest. In het studiejaar 1980-1981 werd een vergelijkend onderzoek uitgevoerd waarin de helft van de tweedejaars studenten tandheelkunde behandelingsplannen leerde opstellen met behulp van het probleemoplossingsmodel en de andere helft met behulp van de traditionele probleemoplossingscyclus. De belangrijkste conclusie was, dat het probleemoplossingsmodel op zijn minst gelijkwaardig was aan de probleemoplossingscyclus (Verdonschot, 1984).



Figuur 2.1: Het probleemoplossingsmodel (Verdonschot, 1982).

In dit proefschrift zal niet worden ingegaan op de vraag naar de effectiviteit van beide oplossingsmethoden. Ze zijn hier slechts besproken met het oog op de in par. 2.3 en 2.4 aan de orde komende methoden voor het vaststellen van probleemoplosvaardigheid bij tandheelkunde-studenten.

2.3 Vaststellen van probleemoplosvaardigheid met behulp van papieren patiënt problemen

2.3.1 Inleiding

Het gebruik van heuristische methoden bij het oplossen van complexe problemen is geen garantie voor het bereiken van acceptabele oplossingen. Geconcretiseerd voor tandheelkundige problemen betekent dit, dat noch het gebruik van de probleemoplossingscyclus noch het gebruik van het probleemoplossingsmodel gegarandeerd zal kunnen leiden tot aanvaardbare behandelingsplannen.
Anders dan bij dié problemen die alleen door toepassing van een algorithme zijn op te lossen en waarbij slechts onderscheid gemaakt kan worden tussen "oplosvaardig" (het algorithme is aan de oplosser bekend) en "niet oplosvaardig" (het algorithme is niet aan de oplosser bekend), moet in de tandheelkunde het begrip "oplosvaardigheid" als een continuum opgevat worden. Tussen de extremen "goede oplosser" en "slechte oplosser" kunnen talloze kwaliteitsonderscheidingen gedacht worden, die aangeven in hoeverre een persoon in staat is om een behandelingsplan op te stellen, aan de hand waarvan de tandheelkundige problemen van de patiënt zo goed mogelijk worden opgelost. "Zo goed mogelijk" moet daarbij geïnterpreteerd worden in termen van Marshall's (1983) definitie van probleemoplosvaardigheid (zie par. 1.3), wat neerkomt op een oplossing die adequaat geacht wordt voor de tandheelkunde terwijl zo economisch mogelijk wordt omgesprongen met de tijd, de kosten en het ongemak voor de patiënt. Tussen beide extreme schaalpunten in ligt ergens het punt dat "voldoende probleemoplosvaardigheid" representeert. Door opgestelde behandelingsplannen te beoordelen wordt getracht vast te stellen wie aan dit criterium voldoet. In de Nijmeegse Subfaculteit Tandheelkunde worden op grond van dergelijke beoordelingen beslissingen genomen over de toelating van studenten tot patiëntbehandeling. In de volgende paragraaf wordt besproken waarom getwijfeld moet worden aan de geldigheid van deze beslissingen.

2.3.2 Tekortkomingen van het "papieren patiënt probleem" als methode voor het vaststellen van probleemoplosvaardigheid

Een Papieren Patiënt Probleem (PPP) is een schriftelijke weergave van de (vooral) tandheelkundige problemen van een om hulp vragende patiënt. De verstrekte informatie bevat naast informatie van algemene aard (bijvoorbeeld: gezondheid, voeding, hygiëne) ook resultaten uit verricht onderzoek (bijvoorbeeld: intra- en extraoraal onderzoek). Op grond van die informatie moeten studenten de

aanwezige problemen identificeren en er een oplossing voor geven in de vorm van een opgesteld behandelingsplan. In het tweedejaars blok "klinische tandheelkunde conserverend I" oefenen studenten zich in het opstellen van behandelingsplannen voor papieren patiënten. De uitslag van een toets, bestaande uit het oplossen van één PPP, is bepalend voor de toelating tot de klinische patiëntbehandeling. Aan de geldigheid van deze selectie wordt om de volgende redenen getwijfeld:

1. weinig aandacht voor het oplosproces
   Ondanks het feit dat studenten speciale formulieren moeten gebruiken, waarop de fasen van de probleemoplossingscyclus in de juiste volgorde zijn afgedrukt, geeft een opgesteld behandelingsplan weinig informatie over het oplossingsproces dat er aan ten grondslag heeft gelegen. Zo kan uit het behandelingsplan bijvoorbeeld niet goed worden afgeleid of de probleemoplosser efficiënt te werk is gegaan. Efficiëntie is, zoals in par. 1.3 verduidelijkt werd, een erg belangrijke eis voor probleemoplosvaardigheid. Een efficiënte probleemoplosser zal proberen om de problemen van een patiënt met een minimum aan tijd, kosten en ongemak te identificeren en op te lossen. Maar met PPP's kan, voor wat het identificeren van problemen betreft, geen onderscheid gemaakt worden tussen efficiënte en inefficiënte probleemoplossers. Immers, alle informatie die nodig kan zijn voor de probleemidentificatie is in de beschrijving opgenomen en dus onmiddellijk ter beschikking van de probleemoplosser. Belangrijke onderdelen van het diagnostisch denkproces, zoals het verwerven van aanwijzingen (cues) en het genereren en evalueren van hypothesen, worden daardoor buiten de toetsing gehouden (Elstein, Shulman & Sprafka, 1978). Voor het oplossen van PPP's wordt "slechts" een beroep gedaan op de vaardigheid om de juiste verrichtingen te selecteren voor de gesignaleerde problemen en op de organisatorische vaardigheid om de verrichtingen zo efficiënt mogelijk te plannen. Voor een selectie-instrument, op grond waarvan beslissingen worden genomen over de toelating tot klinische patiëntbehandeling, lijkt dit onvoldoende.
2. weinig realistisch
   Het onder punt 1 besproken nadeel van PPP's is ook hier weer aan de orde. Het in werkelijkheid moeizame proces van het verwerven van relevante informatie vormt een schril contrast met het oplossen van PPP's, waarin alle belangrijke informatie cadeau wordt gegeven. De werkelijkheidswaarde van PPP's wordt nog verder beperkt door het ontbreken van de mogelijkheid om de geplande behandelingen uit te voeren en terugkoppeling te krijgen over het resultaat. Natuurlijk is het onmogelijk om op papier psychomotorische activiteiten (in de zin van tandheelkundige verrichtingen) te ontplooien, maar daar gaat de belangstelling dan ook niet naar uit. Belangrijk is dat de student terugkoppeling krijgt over het resultaat van de behandeling, waardoor voor hem duidelijk kan worden of de problemen op de juiste wijze zijn aangepakt. Bovendien wordt de oplosser, in de wetenschap met de gevolgen van genomen beslissingen geconfronteerd te worden, min of meer gedwongen om zeer goed na te denken

over het voorgelegde probleem. De risico-factor die daarmee ingebouwd wordt zal de betrokkenheid van de oplosser bij het probleem vergroten, waardoor de oplossing zelf een beter beeld kan geven van de oplossingscapaciteit van de oplosser. Of, zoals Boekaerts (1983) concludeert in een artikel over probleemoplossen: "Wil men de oplossingscapaciteit in een ecologisch valide setting vaststellen dan moet men bereid zijn om de gebruikelijke testen te vervangen door opgaven en taken die voor de leerling werkelijkheidswaarde hebben."

3. vertraagde en irrelevante terugkoppeling
   Met name als PPP's gebruikt worden voor examendoeleinden, zal de terugkoppeling ernstig vertraagd zijn als gevolg van het feit dat het beoordelen van een groot aantal behandelingsplannen veel tijd in beslag neemt. Met het verstrijken van de tijd wordt het nut van de teruggekoppelde informatie (in de vorm van "kennis van de resultaten") kleiner, omdat de oplosser zich niet meer (precies) de overwegingen voor de geest kan halen die geleid hebben tot het doen van bepaalde keuzes. Bovendien zal de terugkoppeling in veel gevallen beperkt blijven tot het verstrekken van de uitslag: geslaagd of gezakt. Volgens Mc Keachie (1976, geciteerd in Buis, 1978) is deze vorm van terugkoppeling nutteloos omdat er geen informatie verstrekt wordt die door de informatie-ontvanger aangewend kan worden voor het opsporen van hiaten in zijn kennis.
4. subjectieve beoordeling
   Het opstellen van een behandelingsplan op basis van een papieren patiënt probleem (PPP) kan als methode vergeleken worden met het beantwoorden van open vragen. Dit betekent dat de voor- en nadelen van de open vraagvorm ook van toepassing zijn op PPP's. Een belangrijk voordeel betreft de vrijheid die de geëxamineerde geboden wordt om de oplossing van het probleem in eigen bewoordingen te beschrijven en te motiveren. De geëxamineerde kan al zijn kennis en creativiteit aanwenden om de oplossing tot "zijn" oplossing te maken. Een belangrijk nadeel is een gevolg van deze vrijheid en behelst het subjectieve karakter van de beoordelingen. Zelfs wanneer omschreven is aan welke eisen de oplossing dient te voldoen, is het toch de beoordelaar die moet beslissen of volgens hem aan de criteria wordt voldaan.
   Verdonschot (1980) onderzocht de kwaliteit van beoordelingen van behandelingsplannen en kwam op grond van een analyse daarvan tot de conclusie dat, gezien de grote verschillen tussen beoordelaars, het aanbeveling verdient om met meerdere beoordelaars te werken. De grootste verschillen tussen beoordelaars traden op bij het beoordelen van klinisch onderzoek, de probleemstelling en het eindoordeel. Tot op heden, echter, wordt nog steeds met slechts één beoordelaar gewerkt en worden vooraf geen criteria vastgesteld waaraan oplossingen dienen te voldoen.

Resumerend kan geconcludeerd worden, dat het PPP geen geschikt instrument is voor het vaststellen van probleemoplosvaardigheid en dat het nemen van beslissingen over de toelating tot de

klinische patiëntbehandeling, op grond van prestaties op een PPP, afgeraden moet worden.
In par. 2.4 wordt beargumenteerd dat de gesignaleerde problemen met PPP's geheel of gedeeltelijk kunnen worden weggenomen door gebruik te maken van patiënt management problemen (PMP's). Verder wordt in die paragraaf ingegaan op de constructie van twee tandheelkundige PMP's.

2.4 Het vaststellen van probleemoplosvaardigheid met behulp van tandheelkundige patiënt management problemen (PMP's)

2.4.1 Argumenten voor het gebruik van PMP's

De argumenten voor het gebruik van PMP's worden ontleend aan de mogelijkheden die zij bieden om de hiervoor genoemde bezwaren tegen PPP's te ondervangen. Gelet op de volgorde waarin die problemen in par. 2.3.2 behandeld zijn, worden hieronder enkele eigenschappen van PMP's besproken die daarvoor een oplossing kunnen bieden.

1. aandacht voor het oplosproces
   PMP's zijn zodanig geconstrueerd dat oplossers op basis van de summiere informatie die in de openingsscène aan hen verstrekt wordt, beslissingen moeten nemen over te ontplooien activiteiten om een diagnose te kunnen stellen. Als gebruik gemaakt wordt van de onzichtbare druktechniek (zie par. 1.5.3.1) voor het verstrekken van terugkoppeling over de genomen beslissingen, dan kan de oplosroute achteraf vrij behoorlijk gereconstrueerd worden, waarmee inzicht verkregen kan worden in de efficiëntie van het oplosproces. "Vrij behoorlijk" omdat bij vergissingen (verkeerde informatie opgevraagd of verkeerde beslissing genomen) meerdere responsen (de terugkoppeling) in dezelfde sectie ontwikkeld kunnen zijn, zonder dat de volgorde waarin dit gebeurd is achterhaald kan worden. Hetzelfde probleem kan zich voordoen bij secties waarin meerdere opties ontwikkeld mogen worden. In hoofdstuk IV komt dit probleem opnieuw aan de orde bij de bespreking van de voordelen van gecomputeriseerde patiënt management problemen (CPMP's).
2. nauwkeurige benadering van de werkelijkheid
   In de openingsscene van een PMP wordt behalve essentiële informatie over de patiënt tevens een beschrijving opgenomen van de rol die de oplosser moet aannemen, van zijn taak en van de omstandigheden waaronder gewerkt moet worden. Anders dan bij een PPP moet de informatie, op grond waarvan een diagnose gesteld kan worden, door de oplosser zelf verzameld worden. Bijvoorbeeld door het stellen van vragen aan de patiënt, door het verrichten van tandheelkundig onderzoek, door het uitvoeren van laboratoriumonderzoek, enz. De informatie (antwoorden van de patiënt, resultaten uit onderzoek) die verstrekt wordt, elke keer als een keuze voor een activiteit gemaakt is, kan sturend werken voor het kiezen van de volgende activiteit. Anders gezegd: PMP's onderscheiden zich van PPP's in het initiëren van een diagnostisch denkproces, bestaande uit:

het verzamelen van informatie, het formuleren van hypotheses, het interpreteren van informatie en het evalueren van hypotheses.
Een ander punt dat bijdraagt aan een zo goed mogelijke nabootsing van de werkelijkheid is de mogelijkheid om geplande verrichtingen "uit te voeren". De aanhalingstekens geven aan dat dit niet al te letterlijk opgevat moet worden. Immers, op papier kunnen geen tandheelkundige verrichtingen worden uitgevoerd. Wat bedoeld wordt, is dat de oplosser op een bepaald moment in het oplosproces kan aangeven dat hij tot behandeling zou willen overgaan. Het grote voordeel van deze mogelijkheid is, dat aan de oplosser terugkoppeling gegeven kan worden over de gevolgen van deze beslissing. De betere benadering van de werkelijkheid kan bij oplossers een grotere betrokkenheid bij het probleem bewerkstelligen en daardoor als relevanter beschouwd worden. Zowel McGuire (1976) als Boekaerts (1983) denken dat dit de motivatie van de probleemoplosser positief kan beïnvloeden, waardoor meer vertrouwen gesteld kan worden in de uitslag van de toets.

3. onmiddellijke terugkoppeling
Elke keer als in een PMP een keuze wordt gemaakt voor een activiteit wordt onmiddellijke terugkoppeling verstrekt aan de oplosser. Zoals in par. 1.5.3.1 werd besproken kan de terugkoppeling bestaan uit het geven van informatie over het al dan niet juist zijn van de ondernomen activiteit of uit het verstrekken van informatie die in een werkelijke situatie ook verkregen zou zijn bij het uitvoeren van een bepaalde activiteit. Het feit dat de terugkoppeling "onmiddellijk" is, heeft als voordeel dat de oplosser dikwijls nog in staat zal zijn om zich de argumenten te herinneren die bij hem geleid hebben tot het doen van bepaalde keuzes. In geval van een onjuiste keuze of het krijgen van irrelevante informatie bestaat daardoor de mogelijkheid om na te gaan of er sprake is geweest van een denkfout en zo ja, hoe de fout voorkomen had kunnen worden. PMP's kunnen op deze manier het aanpak-gedrag van oplossers beïnvloeden.

4. objectieve beoordeling
De beoordeling van PMP's kan machinaal geschieden, hetgeen betekent dat ze objectief scoorbaar zijn. Voor elke keuze die in het PMP gedaan kan worden is van tevoren bepaald hoe die gewaardeerd moet worden. Van beoordelen is dus geen sprake; er wordt slechts nagegaan welke keuzes gemaakt zijn. Dit betekent overigens niet dat de scoring van PMP's een eenvoudige zaak is (zie par. 1.5.3.2).

2.4.2 De constructie van twee tandheelkundige PMP's

2.4.2.1 Inleiding

In par. 2.3.2 werden enkele tekortkomingen gesignaleerd van het PPP als methode voor het vaststellen van probleemoplosvaardigheid. In par. 2.4.1 werd beweerd dat genoemde tekortkomingen

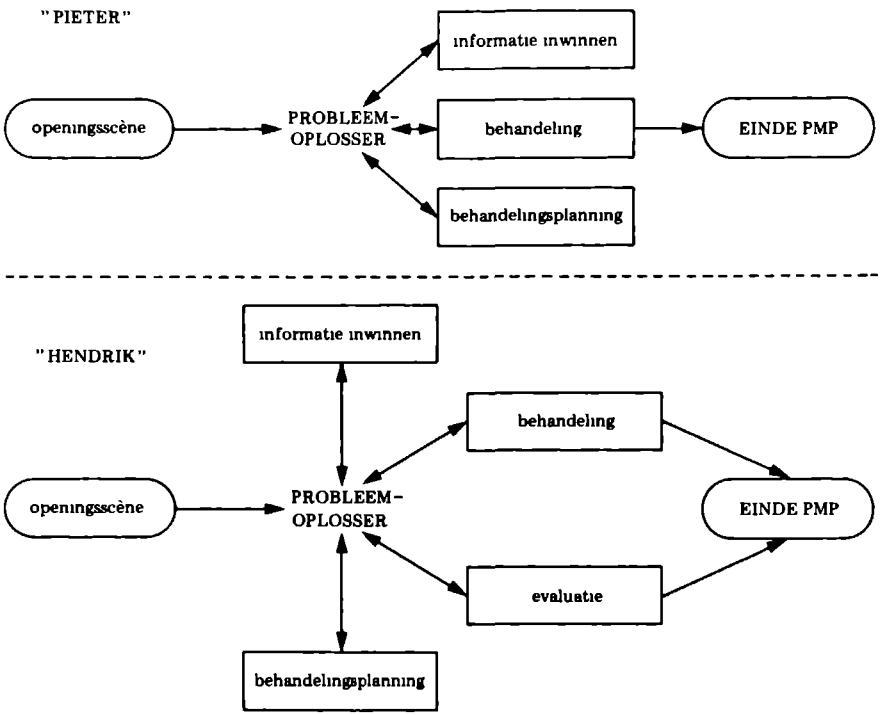
opgeheven zouden kunnen worden door de probleemoplosvaardigheid vast te stellen met behulp van PMP's. Om na te kunnen gaan of PMP's inderdaad beter voor dit doel zijn toegerust werd een studie uitgevoerd, waarin de prestaties van derde-, vierde- en vijfdejaars studenten Tandheelkunde op twee PMP's vergeleken werden met de prestaties op twee (inhoudelijk identieke) PPP's. Over de opzet en de resultaten van deze studie wordt gerapporteerd in hoofdstuk III. In de resterende paragrafen van dit hoofdstuk wordt besproken wat de eigenschappen zijn van de twee geconstrueerde PMP's.

2.4.2.2 Structuur van de PMP's

Volgens de indeling die Harden (1983) hanteert voor PMP-varianten (zie par. 1.5.3.1), zijn de geconstrueerde PMP's te karakteriseren als "open" problemen. Dat wil zeggen dat de oplosser zelf bepaalt welke activiteiten (bijvoorbeeld: opvragen van informatie, interpreteren van informatie, kiezen van een therapie, enz.) ontplooid worden én in welke volgorde. De volgorde zal vooral afhankelijk zijn van de terugkoppeling naar aanleiding van de eerder ondernomen activiteit(en). De keuze voor deze variant is ingegeven door de overweging dat het oplosproces een zo natuurgetrouw mogelijke afspiegeling moet zijn van de werkelijkheid. Echter, om te voorkomen dat oplossers door een vergissing of verkeerde interpretatie beslissingen nemen die leiden tot zeer vroegtijdige beëindiging van het PMP, worden bepaalde beslissingen geblokkeerd door de respons: "kies een andere optie uit deze sectie". Vandaar dat de structuur van de geconstrueerde PMP's nauwkeuriger omschreven is met McGuire's (1976) term: "modified free-branching technique" (zie par. 1.5.3.1).

De vervaardigde PMP's bestaan uit grote aantallen items, die geordend zijn in secties. Onder een item wordt in dit verband verstaan: een keuzemogelijkheid (optie) én de daarbij behorende terugkoppeling (respons). De secties zijn op hun beurt geordend naar de activiteiten die de oplosser ontplooit bij het oplossen van het PMP. In figuur 2.2 wordt voor beide PMP's een overzicht gegeven van de globale structuur. De PMP's worden in het vervolg aangeduid met "Pieter" en "Hendrik", de voornamen van de patiënten in deze simulaties. "Pieter" is een eenvoudiger probleem dan "Hendrik" (in par. 2.4.2.3 wordt uiteengezet welke factoren bepalend zijn geweest voor de moeilijkheidsgraad). Verder verschillen de PMP's van elkaar door het ontbreken van de mogelijkheid bij "Pieter" om terugkoppeling te krijgen over de kwaliteit van het uitgevoerde behandelingsplan. Oplossers van dit PMP krijgen, als ze te kennen hebben gegeven de simulatie te willen beëindigen, alleen informatie over de juistheid van het aantal geïdentificeerde problemen. Beide problemen vangen aan met een openingsscène waarin de probleemoplosser noodzakelijke informatie krijgt over de patiënt, de aan te nemen rol, de opdracht en de omstandigheden waaronder gewerkt moet worden. Daarna is de oplosser in principe vrij om te beslissen welke activiteiten ondernomen worden en in welke volgorde. In de schema's van figuur 2.2 wordt dit aangegeven

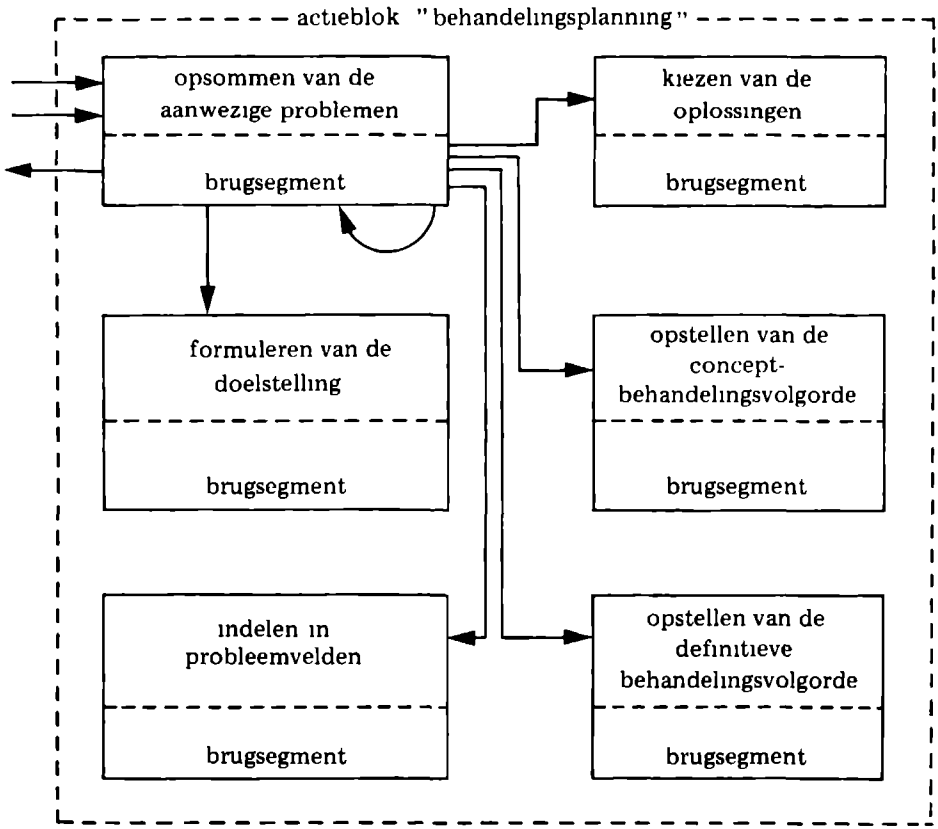
door de bidirectionele pijlen tussen de probleemoplosser en de, tot "actieblokken" gegroepeerde, secties.
Om terugkoppeling te verstrekken aan de oplosser is gekozen voor de "latent image printing"-techniek. In de eerste plaats omdat de PMP's daardoor bruikbaar zijn voor zowel toets- als oefendoeleinden. Als PMP's gebruikt worden om te oefenen is de wijze waarop terugkoppeling verstrekt wordt niet zo belangrijk, zo lang de probleemoplosser maar zelfstandig de benodigde informatie kan verwerven (in par. 1.5.3.1 zijn diverse manieren genoemd waarop dit kan geschieden). Maar, als de PMP's gebruikt worden voor toetsdoeleinden, dan is het noodzakelijk om de terugkoppeling op zodanige wijze te verbergen dat deze alleen verstrekt wordt als de probleemoplosser daar recht op heeft. De onzichtbare druktechniek staat hier borg voor. In de tweede plaats omdat de onzichtbare druktechniek het mogelijk maakt om achteraf oplosprocessen te reconstrueren.



Figuur 2.2: Globale structuur van de vervaardigde PMP's.

Het actieblok "behandelingsplanning" bevat een groot aantal secties, die samen het in par. 2.2 besproken probleemoplossingsmodel representeren. De oplosser wordt de keuze gelaten tussen het oplossen van alle aanwezige problemen tegelijk en het oplossen van de aanwezige problemen in volgorde van urgentie. Dat wil zeggen:

eerst de onmiddellijke problemen, dan de microbiële problemen, vervolgens de functieproblemen en tenslotte de onderhoudsproblemen. Ongeacht de keuze voor een gefaseerde of niet-gefaseerde behandeling, bestaat voor de oplosser de mogelijkheid om op gestructureerde wijze een behandelingsplan op te stellen. Namelijk door in de brugsegmenten steeds dié optie te kiezen die verwijst naar een sectie waarin activiteiten worden ontplooid, die volgens het probleemoplossingsmodel in die bepaalde fase van het oplosproces ontplooid zouden moeten worden. Onder een brugsegment wordt dát gedeelte van een sectie verstaan, waarin de oplosser naar een volgende sectie verwezen wordt (bij lineaire en vertakte problemen) of waarin de oplosser een keuze kan maken uit een aantal activiteiten, om op grond van die keuze doorgestuurd te worden naar een andere sectie (bij vertakte en open problemen). In figuur 2.3 is schematisch weergegeven hoe het probleemoplossingsmodel in de structuur van het PMP verwerkt is. Voor de duidelijkheid is in dit schema alleen de opbouw gegeven van het actieblok "behandelingsplanning". Eveneens ter bevordering van de duidelijkheid zijn in het schema alleen de verbindingen getekend tussen de sectie "opsommen van de aanwezige problemen" en alle andere secties.



Figuur 2.3: Schematische weergave van de integratie van het probleemoplossingsmodel in de structuur van een PMP.

Met andere woorden: een probleemoplosser kan vanuit een bepaalde sectie in het actieblok "behandelingsplanning" naar elke andere sectie gaan in dit actieblok. De keuze hiervoor maakt de probleemoplosser in het brugsegment (onder de stippellijn) van een sectie. Het schema laat zien dat oplossers via twee routes in het actieblok "behandelingsplanning" kunnen komen. Oplossers starten in dit actieblok altijd in het optiesegment (boven de stippellijn) van de sectie "opsommen van de aanwezige problemen". Nadat ze de instructies in dit optiesegment hebben opgevolgd worden ze doorverwezen naar het brugsegment van de betreffende sectie. Daar moeten ze een keuze maken uit een aantal activiteiten. De oplosser maakt zijn keuze duidelijk door de met onzichtbare inkt gedrukte respons achter de optie van zijn keuze te ontwikkelen. De dan zichtbare tekst vertelt de oplosser hoe het oplosproces vervolgt moet worden. De meeste instructies betreffen verwijzingen naar de optiesegmenten uit het actieblok "behandelingsplanning". Verder is er altijd een instructie die verwijst naar een sectie die geen deel uitmaakt van het actieblok "behandelingsplanning" en een instructie die als volgt luidt: "kies een andere optie uit deze sectie". In het schema is deze laatste instructie gerepresenteerd door de in het brugsegment terugkerende pijl.

2.4.2.3 De moeilijkheidsgraad van de vervaardigde PMP's

Volgens De Jong en Ferguson-Hessler (1982) wordt de moeilijkheidsgraad van problemen vooral bepaald door de volgende factoren:
a. het kennisrepertoire van de oplosser;
b. het "aanpakgedrag" van de oplosser;
c. de eigenschappen van het probleem.

ad. a Kennisrepertoire van de oplosser

Zonder relevante kennis kunnen geen problemen worden opgelost. De omvang van de relevante kennis bepaalt voor een deel de moeilijkheidsgraad van het probleem. De omvang van de kennis is wellicht af te leiden uit de vorderingen van de student. Al te optimistische verwachtingen hierover zijn echter misplaatst, gezien het gebruik van voornamelijk normgeoriënteerde toetsen voor de cognitieve blokken. Normgeoriënteerde toetsen drukken, in tegenstelling tot criteriumgeoriënteerde toetsen, de testprestatie uit in een relatieve maat. Het hangt van de prestaties van de gehele groep af of de kennis van een individu als voldoende of onvoldoende wordt bestempeld. Criteriumgeoriënteerde toetsen leveren wél scores op die aangeven welke positie de geëxamineerde inneemt op een continuum dat loopt van "geen kennis" tot "perfecte kennis" (Millman, 1974).
Vooralsnog moet dus geconstateerd worden dat onvoldoende bekend is over de kennis waarover tandheelkunde-studenten kunnen beschikken bij het oplossen van problemen, om dit gegeven van invloed te kunnen laten zijn op de aanduiding van de moeilijkheidsgraad van de vervaardigde PMP's.

ad. b Het aanpakgedrag van de oplosser

De wijze waarop probleemoplossers een probleem benaderen is onder andere afhankelijk van de cognitieve stijl. Een cognitieve stijl is een formele procesvariabele, die van persoon tot persoon verschilt en betrekking heeft op de wijze waarop informatie door de persoon wordt georganiseerd (Drenth, 1973). Vaags (1975) bestudeerde diverse onderzoeken naar de invloed van een cognitieve stijl op het aanpakgedrag van probleemoplossers en concludeerde dat er weinig aanwijzingen waren voor een verband. Maar zelfs in het geval van een bewezen verband zou dit van weinig waarde zijn geweest voor het bepalen van de moeilijkheidsgraad van de PMP's. Immers, er is geen informatie beschikbaar over de cognitieve stijl van de studenten voor wie de PMP's bestemd zijn.

ad. c De eigenschappen van het probleem

Niet in de laatste plaats wordt de moeilijkheidsgraad van een probleem bepaald door kenmerken/eigenschappen van het probleem zelf; bijvoorbeeld de structuur. Een "open" PMP kan als moeilijker worden ervaren dan een "lineair" PMP, omdat in het eerstgenoemde type probleem veel meer aan het eigen inzicht van de probleemoplosser wordt overgelaten. Dit criterium is echter niet hanteerbaar als de moeilijkheidsgraad bepaald moet worden van problemen met gelijke structuur. Een voor de hand liggende en bruikbare indicator voor de moeilijkheidsgraad is de complexiteit van het probleem. De complexiteit wordt groter als:
- de hoeveelheid kennis, benodigd voor het oplossen van het probleem, toeneemt;
- het aantal deelproblemen toeneemt.

De hoeveelheid benodigde kennis kan geoperationaliseerd worden door:
- het aantal doelstellingen te bepalen dat bereikt moet zijn door probleemoplossers die het PMP goed oplossen;
- het niveau te omschrijven van die doelstellingen. Bijvoorbeeld door aan te geven voor welke jaargroep de doelstellingen bedoeld zijn.

Bij de PMP's "Pieter" en "Hendrik" is het onderscheid in moeilijkheidsgraad aangebracht door het ene probleem complexer te maken dan het andere. In tabel 2.2 is aangegeven op welke wijze getracht is dit te bereiken.

Tabel 2.2: Definiëring van de moeilijkheidsgraad van de vervaardigde PMP's.

criteria	Pieter	Hendrik
aantal doelstellingen	7	8
niveau der doelstellingen	2de jaars stof	3de jaars stof
aantal deelproblemen	13	19

2.5 Discussie

In dit hoofdstuk is aandacht besteed aan een tweetal nieuwe ontwikkelingen op het terrein van het oplossen van tandheelkundige problemen. In de eerste plaats aan Verdonschot's probleemoplossingsmodel: een combinatie van heuristische methoden die in een voorgeschreven volgorde moeten worden toegepast op een patiëntprobleem. In de tweede plaats aan de constructie van twee tandheelkundige PMP's die, naar verwachting, beter geschikt zijn voor het vaststellen van probleemoplosvaardigheid dan de traditionele papieren patiënt problemen. De integratie van het probleemoplossingsmodel in de vervaardigde PMP's biedt het voordeel dat het oefenen met het probleemoplossingsmodel in een realistische "setting" kan plaatsvinden. Verwacht mag worden dat het voor studenten daardoor eenvoudiger en vanzelfsprekender wordt om het model te gebruiken als behandelingsplannen moeten worden opgesteld voor echte patiënten. Primair echter zijn de PMP's ontwikkeld uit onvrede met de wijze waarop tot op heden probleemoplosvaardigheid wordt vastgesteld. In par. 2.3.2 zijn vier argumenten genoemd ter verdediging van het standpunt dat PPP's niet geschikt zijn om te discrimineren tussen studenten die probleemoplosvaardig zijn en studenten die dat niet zijn. Dat PMP's daarvoor beter geschikt zijn is beargumenteerd in par. 2.4.1. In hoofdstuk III wordt een studie besproken, die onder meer tot doel had na te gaan of genoemde voordelen van PMP's ook tot uiting komen in een toename van de validiteit.

III EEN STUDIE NAAR DE VALIDITEIT VAN DE GECONSTRUEERDE PMP'S

G.J.J.M. STRAETMANS
E.H.A.M. VERDONSCHOT

3.1 Inleiding

In par. 1.5.3.2 is, bij een bespreking van enkele nadelen van PMP's (patiënt management problemen), het vraagstuk van de validiteit van deze instrumenten summier aan de orde geweest. De resultaten uit validiteitsonderzoek van de daar aangehaalde studies waren zeer uiteenlopend, waardoor algemene uitspraken over de validiteit van PMP's achterwege moesten blijven. Overigens zijn uitspraken over dé validiteit van PMP's niet mogelijk, omdat deze niet als zodanig bestaat. "Validiteit" is een overkoepelende term voor een aantal soorten validiteit die elk de kwaliteit van een instrument op andere wijze benaderen. In par. 3.2 zullen de belangrijkste soorten validiteit besproken worden en tevens de vraagstellingen die geformuleerd zijn voor het onderzoek daarnaar. In par. 3.3 volgt een beschrijving van de opzet van het onderzoek en in par. 3.4 worden per vraagstelling de resultaten gepresenteerd. Par. 3.5 is gewijd aan een bespreking van de resultaten. Het hoofdstuk wordt afgesloten met een opsomming van de conclusies en enkele aanbevelingen (par. 3.6).

3.2 Probleemstelling

Het belangrijkste doel van het in dit hoofdstuk beschreven onderzoek is na te gaan welke waarde gehecht kan worden aan de oplossingen van de PMP's "Pieter" en "Hendrik", de twee tandheelkundige PMP's waarvan de eigenschappen besproken werden in par. 2.4.2.2 en 2.4.2.3. Wat betekent het bijvoorbeeld, als een respondent een hoge score behaalt op een van deze PMP's? Kan een dergelijke hoge score representatief geacht worden voor de probleemoplosvaardigheid van die respondent met betrekking tot tandheelkundige, klinische problemen? Beknopt geformuleerd luidt de vraagstelling: "Hoe valide zijn de geconstrueerde PMP's?"

Zoals gezegd in de inleiding, wordt de term "validiteit" gebruikt als aanduiding voor verschillende soorten validiteit, die dit begrip elk op andere wijze benaderen. In een onderzoek naar de validiteit van medische PMP's bestudeerden Page en Fielding (1980) de inhoudsvaliditeit, de constructvaliditeit en de criteriumvaliditeit. Hieronder worden deze soorten besproken. Tevens wordt aangegeven of, en zo ja hoe, de verschillende soorten validiteit vastgesteld kunnen worden voor de vervaardigde PMP's "Pieter" en "Hendrik".

a. inhoudsvaliditeit

Volgens Drenth (1973) kan van de inhoudsvaliditeit een schatting worden verkregen door experts te laten beoordelen hoezeer de inhoud van de test een hele klasse situaties, kennisinhouden of vaardigheden representeert waarover conclusies moeten worden getrokken. Inhoudsvaliditeit heeft betrekking op de generaliseerbaarheid van de testresultaten. In hoeverre mag vanuit de prestatie op een tandheelkundig PMP gegeneraliseerd worden naar probleemoplosvaardigheid met betrekking tot tandheelkundige, klinische problemen in het algemeen? Een belangrijk bezwaar van dit type validiteit is dat empirische verificatie vaak niet mogelijk is, waardoor aan het consensus-oordeel van experts gauw te veel waarde wordt gehecht.
Bij de constructie van de twee tandheelkundige PMP's zijn maatregelen genomen om een voldoende mate van inhoudsvaliditeit te waarborgen. Zoals in par. 2.4.2.3 is uiteengezet zijn de PMP's "Pieter" en "Hendrik" gebaseerd op een verschillend aantal leerdoelen, die representatief geacht worden voor de leerstof uit bepaalde studiejaren. De PMP's "Pieter" en "Hendrik" representeren een grotere groep van leerdoelen uit respectievelijk het tweede en derde studiejaar. Er wordt van uitgegaan dat genoemde maatregelen hebben geleid tot PMP's die voldoende inhoudsvaliditeit bezitten. In het onderhavige onderzoek is derhalve niet geprobeerd om kwantitatieve evidentie te verkrijgen voor dit type validiteit.

b. constructvaliditeit

Constructvaliditeit wordt geëvalueerd door te onderzoeken welke psychologische kwaliteit(en) een test meet. Cronbach en Meehl (1955, geciteerd in De Zeeuw, 1978) omschrijven de term "construct" als volgt: "Een of ander gepostuleerd attribuut, waarvan men veronderstelt dat het gereflecteerd wordt in de testprestatie." Onderzoek naar de constructvaliditeit verloopt veelal als min of meer gericht exploratief onderzoek. De Zeeuw (1978) omschrijft deze onderzoekswerkzaamheden als volgt: "Er is tenminste een doelstelling en er is een idee over de verwerkelijking daarvan door een test, zodat de items een bepaalde vorm of inhoud hebben. Zij zijn zodanig gekozen, of samengesteld, dat zij vermoedelijk in relatie staan met hetgeen onderzocht moet worden en dat omschreven is in het testbegrip, als verbale definiëring van een psychische eigenschap of trek, ook aangeduid met "attribuut". Gesteld dat een onderzoeker het "inprentingsvermogen" wil onderzoeken, weergegeven door I, en dat hij een non-verbale test X daarvoor heeft ontworpen, welke een voldoende mate van betrouwbaarheid bezit. Hij heeft daarmee vanzelfsprekend nog geen enkele zekerheid dat deze test ook werkelijk het bedoelde meet; ja, hij weet niet eens precies, wat het begrip "inprentingsvermogen" omvat. Hij veronderstelt echter dat hij empirisch tussen X en andere representanten van I verbanden kan aantonen, daarbij al dan niet uitgaande van een (voorlopig) theoretisch concept omtrent "inprenting". De onderzoeker stelt toetsbare hypothesen op betreffende die ver-

banden. Bijvoorbeeld: inprenting houdt sterk verband met concentratie; test X zal derhalve positief correleren met de erkende concentratietest C op het significantieniveau van tenminste a."
In het onderhavige onderzoek gaat het over het construct "probleemoplosvaardigheid". Aangezien het oplossen van tandheelkundige problemen een ingewikkeld samenspel is van cognitieve, affectieve en psychomotorische vaardigheden, is het moeilijk om daaruit één aspect te kiezen dat geacht wordt indicatief te zijn voor probleemoplosvaardigheid. Daarom is besloten om de constructvaliditeit van de vervaardigde PMP's, in navolging van soortgelijke studies door Page en Fielding (1980) (zie par. 1.5.3.2) en Newble, Hoare en Baxter (1982), te evalueren door het toetsen van de hypothese, dat vijfdejaars studenten hogere scores zullen behalen op de PMP's dan de vierde- en derdejaars studenten en dat de vierdejaars beter zullen presteren dan de derdejaars studenten. Deze hypothese is gebaseerd op het vermoeden dat de omvang van de voor het oplossen van tandheelkundige problemen relevante kennis, vaardigheden en ervaring, nauw gerelateerd is aan de studievordering.

c. criteriumvaliditeit

Criteriumvaliditeit wordt meestal geschat op basis van een of andere correlatiestudie. De essentie is dat de onderzoeker probeert om een ander (liefst meer direct) instrument te vinden waarmee de competenties, die het nieuw ontwikkelde instrument beoogt te meten, vastgesteld kunnen worden. De scores op beide instrumenten worden vervolgens gecorreleerd. De moeilijkheid bij dit type validiteitsonderzoek ligt in het vinden van een geschikt criterium. Conventionele objectieve tests representeren eerder minder dan meer directe maten van de complexe vaardigheden die simulaties beogen vast te stellen. Niet-objectieve tests, zoals bijvoorbeeld "on-the-job-tests", zijn vaak niet geschikt als gevolg van de lage inter-beoordelaarsbetrouwbaarheid en als gevolg van het feit dat ze vaak gebaseerd zijn op relatief weinig waarnemingen (McGuire, 1976).
Voor het onderhavige onderzoek kon geen geschikt criterium gevonden worden in de zin van scores op een ander instrument dat probleemoplosvaardigheid beoogt te meten. In plaats daarvan werden de scores op de PMP's gecorreleerd met scores op een aantal cognitieve toetsen, die studenten in de loop van hun studie hadden afgelegd. De argumentatie hiervoor werd ontleend aan de resultaten van recent fundamenteel onderzoek op het gebied van het probleemoplossen. Greeno (1980) vat deze resultaten samen in de conclusie dat het onderscheid tussen "probleemoplosvaardigheid" en "kennis" lang niet zo scherp is als lange tijd is aangenomen: "Modern research in cognitive psychology and artificial intelligence shows that knowledge is at the basis of all problem solving. Our educational task is to discover what knowledge is needed for a class of tasks and then to discover how to communicate that knowledge effectively to our students."
De verwachting is dat studenten met hogere cijfers voor de cognitieve toetsen beter zullen presteren op de PMP's dan studenten met

lagere cijfers.

Tot nu toe is alleen gesproken over validering van de vervaardigde PMP's. Daarnaast, echter, is het van belang om te onderzoeken of PMP's beter geschikt zijn voor het meten van probleemoplosvaardigheid dan de tot op heden gebruikte papieren patiënt problemen (PPP's). Ten behoeve van deze vergelijking werden twee PPP's vervaardigd die inhoudelijk identiek zijn aan de geconstrueerde PMP's. Alle hiervoor besproken validiteitsstudies zullen eveneens uitgevoerd worden voor de PPP's.

De volgende vraagstellingen worden onderzocht:

1. Is de volgorde waarin PMP's of PPP's worden aangeboden (eerst een eenvoudig, daarna een moeilijker probleem of omgekeerd*) van invloed op de kwaliteit van de oplossingen?
2. Zijn de geïntroduceerde moeilijkheidsgraden (zie paragraaf 2.4.2.3) geldig?
3. Zijn er verschillen tussen de derde-, vierde- en vijfdejaars studenten met betrekking tot de prestaties op PMP's en PPP's? (constructvaliditeit)
4. Zijn er verschillen in oplossingskwaliteit tussen een PMP en PPP van dezelfde moeilijkheidsgraad?
5. Zijn er verschillen tussen de derde-, vierde- en vijfdejaars studenten met betrekking tot de wijze waarop een PMP van een bepaalde moeilijkheidsgraad wordt aangepakt? (constructvaliditeit)
6. Is er sprake van een verband tussen de prestaties op een PMP of PPP en prestaties op cognitieve toetsen? (criteriumvaliditeit)
7. Wat zijn de meningen van tandheelkunde studenten met betrekking tot het oplossen van PMP's en PPP's?

Vraagstelling 1 is van belang in verband met het eventuele optreden van korte-termijn leereffecten, als gevolg van het achtereenvolgens oplossen van twee problemen. Vermoedelijk zal een dergelijk effect zich eerder voordoen bij de onbekende PMP's dan bij de vertrouwde PPP's. Het is niet ondenkbaar dat moeilijkheden die zich voordoen bij de confrontatie met het eerste probleem, niet meer terugkeren als aan het tweede probleem gewerkt wordt. Als de problemen die in tweede instantie worden aangeboden significant beter worden gemaakt, dan moet bij volgende analyses rekening worden gehouden met de "volgorde" als extra variabele. Dit kan van invloed zijn op de keuze van de statistische methoden ten behoeve van de andere vraagstellingen.

*Zie tabel 3.1 voor een schematische weergave van de opzet van het onderzoek.

Vraagstelling 2 onderzoekt of de gedefinieerde moeilijkheidsgraden (zie par. 2.4.2.3) geldig zijn.
Met vraagstelling 3 wordt de constructvaliditeit van de PMP's en PPP's onderzocht.
Vraagstelling 4 onderzoekt of de testmethode (PMP of PPP) van invloed is op de prestaties van de respondenten.
Vraagstelling 5 heeft evenals vraagstelling 3 tot doel om te onderzoeken of de geconstrueerde problemen constructvaliditeit bezitten. Anders dan bij vraagstelling 3 wordt alleen gekeken naar de gevolgde oplosroutes, hetgeen impliceert dat alleen de prestaties op de PMP's bestudeerd kunnen worden.
De criteriumvaliditeit van de PMP's en PPP's wordt onderzocht door middel van vraagstelling 6.
Vraagstelling 7, ten slotte, heeft tot doel om de meningen en ervaringen van de studenten te inventariseren met betrekking tot het oplossen van management problemen. Hun bevindingen zullen gebruikt worden bij de ontwikkeling van nieuwe PMP's.

3.3 Materiaal en methoden

3.3.1 Materiaal

De studie die een antwoord zou moeten geven op de hiervoor geformuleerde vraagstellingen bestond uit de afname van een toets direct na de start van het studiejaar 1983-1984 aan de Subfaculteit Tandheelkunde van de Katholieke Universiteit in Nijmegen. Deze toets, waaraan deelname verplicht was voor alle derde-, vierde- en vijfdejaars studenten, bestond uit het oplossen van twee tandheelkundige management problemen. Van de 236 in aanmerking komende studenten konden er 14 niet meedoen door ziekte of stagewerkzaamheden.

In de studie werd met vier verschillende tandheelkundige management problemen gewerkt: de PMP's "Pieter" en "Hendrik" en twee PPP's, aangeduid met de namen "Petra" en "Johanna". PMP "Pieter" was inhoudelijk identiek aan PPP "Petra" en PMP "Hendrik" had dezelfde inhoud als PPP "Johanna". Voor het gemak zullen de management problemen in het vervolg worden aangeduid met:
- PMP1: het eenvoudige PMP (= Pieter);
- PMP2: het moeilijke PMP (= Hendrik);
- PPP1: het eenvoudige PPP (= Petra);
- PPP2: het moeilijke PPP (= Johanna).

Om de interne structuur van elk probleem hetzelfde te laten zijn werden alle problemen geconstrueerd door één tandarts-medewerker van het Instituut Conserverende Tandheelkunde voor Volwassenen van voornoemde Universiteit. Omdat sommige studenten twee problemen van dezelfde moeilijkheidsgraad kregen aangeboden (tabel 3.1) werd, om ongewenste transfer tegen te gaan, geprobeerd de inhoudelijke overeenkomst zo goed mogelijk te verbergen door het veranderen van de namen waarmee de deelproblemen werden aangeduid.

3.3.2 Methoden

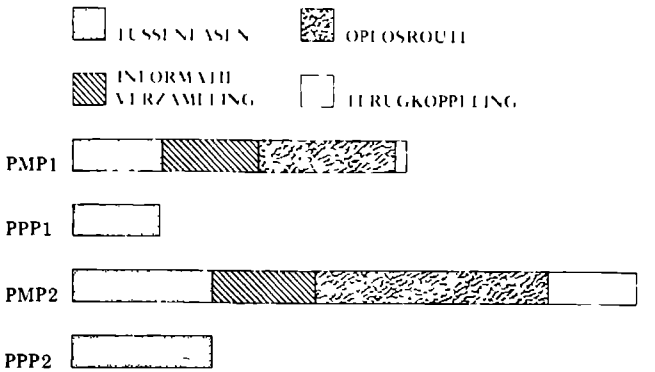
Voor het achtereenvolgens oplossen van twee management problemen kregen studenten drie uur de tijd. De studenten waren op de hoogte van de doelstellingen van de studie waarin ze participeerden, evenals van het feit dat de resultaten geen invloed zouden hebben op beslissingen over studievorderingen. Elke sessie begon met een uiteenzetting over kenmerken en werkwijze van PMP's. Ten einde gedrag uit te lokken waaruit "probleemoplosvaardigheid" zou kunnen blijken, werden studenten geïnstrueerd om de problemen zo goed en zo snel mogelijk op te lossen. Daarbij moesten ze uitgaan van de veronderstelling tandarts te zijn in een algemene praktijk.

Binnen elk studiejaar werden alle studenten op strikt toevallige wijze verdeeld over zes groepen. Deze groepen verschilden van elkaar in de combinatie van aan te bieden management problemen. Tabel 3.1 laat de verschillende combinaties zien.

Tabel 3.1: Voorkomende combinaties van management problemen in elk studiejaar.

	groep					
volgorde	1	2	3	4	5	6
als eerste opgelost	PMP1	PMP2	PMP1	PMP2	PPP1	PPP2
als tweede opgelost	PMP2	PMP1	PPP1	PPP2	PPP2	PPP1

Elke groep in een studiejaar bestond uit elf tot dertien studenten, afhankelijk van het totale aantal studenten in een studiejaar. De ruwe data in deze studie werden verkregen door het (machinaal) scoren van de gemaakte keuzen in de PMP's en van de opgestelde behandelingsplannen voor de PPP's en PMP's. Figuur 3.1 laat zien hoe de ruwe data zijn opgebouwd. De lengte van de staven is indicatief voor de relatieve hoeveelheid informatie die elk management probleem bevat voor de onderscheiden componenten. Uit figuur 3.1 blijkt dat elk opgelost management probleem informatie oplevert over de tussenfasen in het oplosproces. Daarnaast geven opgeloste PMP's informatie over het verzamelen van gegevens, de gevolgde oplosroute en de verstrekte terugkoppeling.



Figuur 3.1: Opbouw van de ruwe data in de verschillende management problemen.

3.3.2.1 Cijferbepalings-systemen voor de PMP's en PPP's

Bij het bepalen van cijfers voor opgeloste PMP's en PPP's werd de component "terugkoppeling" ondergebracht in de component "oplosroute". Eindcijfers voor PMP's werden dus opgebouwd uit drie componenten terwijl het eindcijfer voor een PPP door slechts één component bepaald werd. De gemeenschappelijke component in alle PMP's en PPP's bestond uit de volgende tussenfasen van het probleemoplossingsproces (Zie tabel 2.1 voor een volledige opsomming van de tussenfasen die volgens Verdonschot's probleemoplossingsmodel in het oplosproces te onderscheiden zijn):
1. de probleemidentificatie/probleemformulering;
2. de gekozen oplossingen voor de geïdentificeerde problemen;
3. de definitieve behandelingsvolgorde, bestaande uit:
   - een volgorde van gekozen behandelingsalternatieven;
   - een schatting van de behandelingskosten;
   - een schatting van de behandelingstijd.

Aangezien al deze tussenfasen resulteren in een te beoordelen produkt wordt het toegekende cijfer voor deze component aangeduid met de term "produktcijfer".
Evaluatie van de andere twee componenten ("verzamelen van informatie" en "oplosroute") resulteert in een "procescijfer".
De volledige beoordeling van een PMP is gebaseerd op een evaluatie van alle componenten en wordt uitgedrukt in een "totaalcijfer".

Omdat bij de berekening van de cijfers voor de prestaties op de PMP's en PPP's niet elke component even zwaar mee mocht tellen, werden gewichten bepaald voor de te onderscheiden componenten. Voor dit doel gaven zeventien tandartsen van het Instituut Conserverende Tandheelkunde voor Volwassenen hun mening over het rela-

tieve belang van elke component. Ze drukten het belang van elke component voor de beoordeling van de totale prestatie uit in een percentage. Tabel 3.2 geeft een overzicht van hun gemiddelde beoordelingen.

Tabel 3.2: Procentuele gewichten voor de beoordelingscomponenten van PMP's en PPP's. (N = 17)

	component	gem.	s.d.	gew.
PRODUKT (PMP/PPP)	-tussenfasen			
	*probleemidentificatie	14.1	4.8	14
	*aanvaardbaarheid van de gekozen oplossingen	15.9	4.8	16
	*behandelingsvolgorde	11.3	5.5	11
	*schatting behandelingskosten	8.2	4.0	8
	*schatting behandelingstijd	6.3	3.0	6
PROCES (PMP)	-verzamelen van informatie	20.9	5.9	21
	-oplosroute	23.5	6.5	24

Met uitzondering van de schattingen voor "behandelingskosten", "behandelingstijd" en "behandelingsvolgorde" is er sprake van een grote mate van overeenstemming tussen de ondervraagde tandartsen. Reden waarom de gewichten gebaseerd werden op de gemiddelden. Het cijfer voor een PPP wordt berekend uit de deelcijfers behaald op de vijf genoemde onderdelen van de component "tussenfasen". De gewichten van deze onderdelen tellen op tot een totaal van 55%, zodat het onderdeel "probleemidentificatie" maximaal voor 14/55 deel het cijfer bepaalt. Voor het onderdeel "aanvaardbaarheid van de gekozen oplossingen" is dit maximaal 16/55, enz. De gewichten van de componenten die een rol spelen bij de bepaling van totaalcijfers voor PMP's tellen op tot 100%. Bij een PMP wordt het totaalcijfer dus maximaal voor 14/100 deel bepaald door het deelcijfer behaald op het onderdeel "probleemidentificatie". Het cijfer voor een PPP (= produktcijfer) wordt berekend door de som van alle deelcijfers van de component "tussenfasen" te delen door 5.5. Voor een PMP wordt het totaalcijfer bepaald door de som van deelcijfers van alle componenten te delen door 10. Het produktcijfer voor een PMP wordt berekend door de som van de deelcijfers voor de component "tussenfasen" te delen door 5.5. Het procescijfer voor een PMP wordt bepaald door de som van de deelcijfers voor de componenten "informatieverzameling" en "oplosroute" te delen door 4.5.

Voor elk van de zeven aspecten waarop management problemen beoordeeld kunnen worden, zal hieronder beschreven worden op welke

wijze daarvoor deelcijfers werden berekend.

a. probleemidentificatie

Besloten werd dat een cijfer voor dit onderdeel gebaseerd zou moeten zijn op het aantal geïdentificeerde deelproblemen in het PMP en PPP. Het deelcijfer werd berekend aan de hand van de volgende formule:

$$\frac{\text{aantal geïdentificeerde deelproblemen}}{\text{totaal aantal deelproblemen}} * \text{gewicht}$$

b. aanvaardbaarheid van de gekozen oplossingen

Om een deelcijfer te kunnen berekenen voor dit onderdeel werden alle gekozen alternatieven gescoord op een driepuntsschaal, waarbij een "2" stond voor zeer acceptabel, een "1" voor acceptabel en een "0" voor niet acceptabel. De formule die gebruikt werd om een deelcijfer te berekenen luidt als volgt:

$$\frac{\text{aantal bonuspunten voor aanvaardbaarheid}}{2 * \text{aantal geïdentificeerde problemen}} * \text{gewicht}$$

c. behandelingsvolgorde

Voor elk deelprobleem in een PMP en PPP was door de constructeur bepaald op welk moment dit het beste opgelost zou kunnen worden. Twee strafpunten werden toegekend als op onaanvaardbare wijze van deze ideale behandelingsvolgorde werd afgeweken. Minder ernstige afwijkingen werden bestraft met één strafpunt. Het opstellen van de ideale behandelingsvolgorde leverde twee bonuspunten op. Eén bonuspunt werd toegekend als de volgorde aanvaardbaar was. Als voor een gekozen oplossing de volgorde niet van belang was, dan werden bonus- noch strafpunten toegekend. Het deelcijfer voor de behandelingsvolgorde werd als volgt berekend:

$$\frac{\text{(aantal bonuspunten - aantal strafpunten)}}{\text{maximum aantal bonuspunten}} * \text{gewicht}$$

d. schatting van de behandelingskosten

Zeventien tandartsen van het Instituut Conserverende Tandheelkunde voor Volwassenen werden ondervraagd over de aanvaardbaarheid van schattingen van behandelingskosten. In tabel 3.3 wordt hun gemiddelde mening weergegeven. De cijfers in de tabel zijn percentages die aangeven hoe groot afwijkingen (ten opzichte van

de werkelijke kosten) maximaal mogen zijn om nog met een voldoende gehonoreerd te kunnen worden. Afwijkingen werden lineair getransformeerd naar een honderdpunts-schaal door gebruik te maken van onderstaande formule:

$$\frac{(100 - \text{afwijkingspercentage} * 2)}{100} * \text{gewicht}$$

Tabel 3.3: Aanvaardbaarheid van afwijkingen (in percentages) in de schattingen voor behandelingskosten en behandelingstijd.

	behand.kosten	behand.tijd
max. toelaatbare overschatting	20	30
max. toelaatbare onderschatting	20	25

e. schatting van de behandelingstijd

Het deelcijfer voor de geschatte behandelingstijd werd op soortgelijke wijze berekend als dat voor de behandelingskosten. Maar, zoals uit tabel 3.3 blijkt, is de tolerantie voor een overschatting wat groter dan voor een onderschatting. Vandaar dat er twee formules werden ontwikkeld voor het berekenen van een deelcijfer.

Bij overschatting:

$$\frac{(100 - \text{afwijkingspercentage} * 1.33)}{100} * \text{gewicht}$$

Bij onderschatting:

$$\frac{(100 - \text{afwijkingspercentage} * 1.60)}{100} * \text{gewicht}$$

f. verzamelen van informatie

Om prestaties met betrekking tot het verzamelen van informatie te kunnen beoordelen werd een scoringssysteem van bonus- en strafpunten ingevoerd. De keuze voor een zinvolle (de oplossing dich-

terbij brengende) optie werd beloond met één bonuspunt. De keuze voor een schadelijke (de oplossing belemmerende) optie leverde één strafpunt op. Neutrale (de oplossing niet beïnvloedende) opties leverden geen punten op. De volgende parameters waren van belang voor de bepaling van het deelcijfer:

- het aantal zinvolle opties in de informatieverzamelings-secties van een PMP (A)
- het aantal ontwikkelde, zinvolle opties (B)
- het aantal schadelijke opties in de informatieverzamelings-secties van een PMP (C)
- het aantal ontwikkelde, schadelijke opties (D)

Niet ontwikkelde, zinvolle opties leverden één strafpunt op. Het aantal niet ontwikkelde, zinvolle opties (E) werd berekend door: $E = A - B$.
De formule voor de berekening van het deelcijfer werd gedicteerd door de volgende twee extreme situaties, die allebei moesten leiden tot het deelcijfer 0:

- alle opties zijn ontwikkeld;
- geen enkele optie is ontwikkeld.

Het aantal strafpunten als gevolg van onvolledige informatieverzameling (F) werd berekend door:

$$F = \frac{E * (C + 0.5 (A - C))}{A}$$

Het aantal strafpunten als gevolg van het ontwikkelen van schadelijke opties (G) werd berekend door:

$$G = D * \frac{(C + 0.5 (A - C))}{C}$$

Het deelcijfer voor het verzamelen van informatie, ten slotte, werd berekend met de formule:

$$\frac{(A + C) - (F + G)}{A + C} * \text{gewicht}$$

g. oplosroute

Evaluatie van de oplosroute vond plaats door na te gaan of de tussenfasen in het oplosproces ten opzichte van elkaar een volgorde innamen als voorgeschreven door het probleemoplossingsmodel (zie par. 2.2). Elke tussenfase in de juiste volgorde werd beloond met één bonuspunt. Als een student de instructies negeerde door meer dan één optie te ontwikkelen (in de brugsegmenten mag slechts één keuze gemaakt worden), dan leverde dit strafpunten op. Het deelcijfer voor de oplosroute werd bepaald door de volgende

drie parameters:
- het maximale aantal bonuspunten, gegeven de oplossing (H);
- het aantal verworven bonuspunten (K);
- het aantal verworven strafpunten (L).

De formule voor de berekening van het deelcijfer luidde als volgt:

$$\frac{K - L}{H} * \text{gewicht}$$

Alle in de volgende paragrafen te bespreken analyses zijn uitgevoerd met cijfers, berekend aan de hand van het hiervoor beschreven cijferbepalings-systeem. Aan de basis van dat systeem liggen de ruwe scores die toegekend zijn op grond van ontplooide activiteiten (ontwikkelde responsen, prioriteitsbepalingen, geselecteerde behandelingen) tijdens het oplosproces. Op voorhand zijn gewichten toegekend aan alle mogelijke activiteiten. Al naar gelang het nut ervan voor de oplossing van het probleem, worden ontplooide activiteiten dus verschillend gescoord. Betrouwbaarheid en validiteit van cijfers zijn daardoor direct afhankelijk van de betrouwbaarheid en validiteit van de toegekende gewichten (zie ook par. 1.5.3.2 onder punt 2). Doordat de gewichten werden toegekend door slechts één persoon, kan geen zekerheid worden verkregen over de kwaliteit van de gewichten en de daaruit voortkomende cijfers.

3.3.2.2 Statistische analyses

De in par. 3.2 besproken vraagstellingen werden onderzocht door de verkregen data te analyseren met behulp van geschikte statistische methoden. Alle statistische toetsingen werden uitgevoerd op het vijf procent significantieniveau.

Student's t-test werd gebruikt om eventuele effecten op te sporen, veroorzaakt door de volgorde waarin problemen werden aangeboden (vraagstelling 1). Deze vraagstelling werd als eerste onderzocht omdat een significant volgorde-effect van invloed zou zijn op de keuze van statistische methoden voor de analyse van de andere vraagstellingen.

Multivariate variantie-analyses met herhaalde metingen werden uitgevoerd om de geldigheid te onderzoeken van de gedefinieerde moeilijkheidsgraden van de management problemen (vraagstelling 2). De keuze voor een replicatie-design kwam voort uit de beschikbaarheid van twee scores van elk subject op de responsvariabele (prestatie op een PMP of PPP). De subjecten in deze analyse waren afkomstig uit de groepen 1, 2, 5 en 6 (tabel 3.1).

Vraagstelling 3 (verschillen tussen jaargroepen met betrekking tot de testprestaties) en vraagstelling 4 (verschillen in testprestatie tussen PMP's en PPP's van dezelfde moeilijkheidsgraad)

konden eveneens onderzocht worden door het uitvoeren van multivariate variantie-analyses met herhaalde metingen. Voor de beantwoording van beide vraagstellingen kon gebruik gemaakt worden van de testprestaties van dezelfde subjecten (groep 3 en 4 in tabel 3.1).
Vraagstelling 4 kon bovendien nog onderzocht worden door middel van een normale tweevoudige variantie-analyse, uitgevoerd over de testprestaties van de subjecten in groep 1, 2, 5 en 6 (tabel 3.1). De subjecten in deze groepen waren zodanig verdeeld over de condities "jaargroep" en "testmethode", dat per subject niet meer dan één meting was opgenomen in de uit te voeren analyse.

Aan de hand van frekwentietabellen werd nagegaan of er verschillen bestonden tussen studiejaren met betrekking tot de probleemaanpak (vraagstelling 5). Deze probleemaanpak werd afgeleid uit de relatieve positie die probleemgroepen van uiteenlopende urgentiegraad (zie par. 2.2) ten opzichte van elkaar innamen in de definitieve behandelingsvolgorde.

Om vraagstelling 6 te onderzoeken werden de cijfers (proces- en produktcijfers) voor PMP's en PPP's gecorreleerd met eerder behaalde cijfers op cognitieve toetsen. Voor de selectie van die cognitieve toetsen werden de volgende criteria gehanteerd:
- cijfers dienden tot stand te zijn gekomen op basis van expliciet beschreven score- en transformatieregelingen;
- toetsvorm (bijvoorbeeld meerkeuzevragen of open vragen) en genoemde score- en transformatieregelingen moesten onveranderd zijn gebleven gedurende vier, aan het studiejaar 1983-1984 voorafgaande studiejaren;
- de toetsen dienden voldoende betrouwbaar te zijn.

Toepassing van deze criteria resulteerde in een selectie van tien uit een totaal van 33 cognitieve blokken. In tabel 3.4 wordt een overzicht gegeven van de geselecteerde cognitieve blokken alsmede van de studiejaren waarin ze met een toets zijn afgesloten door de studenten die in het onderzoek participeerden.
Bij de berekeningen van de Pearson product moment correlatie coëfficiënten tussen de cijfers voor de management problemen en de cijfers voor de geselecteerde cognitieve toetsen, dienden laatstgenoemde cijfers als predictoren. Ten behoeve van de standaardisatie werden alleen cijfers opgenomen die afkomstig waren van eerste toetspogingen. Om die reden werden de prestaties van recidivisten buiten de berekeningen gehouden.

Tabel 3.4: Overzicht van de cognitieve blokken waarvan de toetsen werden aangewend als predictoren in een studie naar de criteriumvaliditeit.

nummer	bloktitel	studiejaar 3	4	5
100	organen en orgaansystemen	81-82	80-81	79-80
103	orale weefsels en structuren	81-82	80-81	79-80
106	parodontium	81-82	80-81	79-80
108	wetenschappelijke scholing	81-82	80-81	79-80
203	morfologie hoofd/halsgebied	82-83	81-82	80-81
206	parodontium	82-83	81-82	80-81
300	organen en orgaansystemen		82-83	81-82
306	parodontium		82-83	81-82
312	restauratie en materialen		82-83	81-82
403	oro-fac. struct. en traumatologie			82-83

De waardering van de deelnemende studenten voor een bepaald management probleem werd direct na voltooiing ervan "gemeten" aan de hand van een vragenlijst (zie bijlage 7). Vraagstelling 7 werd onderzocht door per probleem de gemiddelde score op elk item van de vragenlijst te berekenen. Daarnaast werden verschillen in waardering tussen studiejaren en tussen management problemen getoetst met behulp van Student's t-test.

3.4 Resultaten

Achtereenvolgens zullen per vraagstelling de resultaten van de uitgevoerde analyses gepresenteerd worden.

3.4.1 Is de volgorde waarin PMP's of PPP's worden aangeboden (eerst een eenvoudig, daarna een moeilijker probleem of omgekeerd) van invloed op de kwaliteit van de oplossingen? (vraagstelling 1)

Voor elk management probleem werd nagegaan of er sprake was van prestatieverschillen tussen studenten die het als eerste probleem en studenten die het als tweede probleem hadden opgelost (tabel 3.1). Tabel 3.5 bevat de resultaten van deze vergelijking voor beide PMP's. Om het effect na te kunnen gaan van het beoordelen van het oplosproces ("verzamelen van informatie", "oplosroute") zijn de gemiddelde prestaties op de PMP's uitgedrukt in een produkt- én een totaalcijfer. De gemiddelden liggen over het algemeen dicht rond de caesuur (6.00). Negatieve t-waarden geven aan dat een

bepaald probleem beter is opgelost door de studenten die het als tweede probleem kregen aangeboden. In zes gevallen is het gemiddelde cijfer voor een bepaald PMP dat als eerste probleem is aangeboden, hoger dan het gemiddelde cijfer voor hetzelfde probleem, dat als tweede werd aangeboden. De omgekeerde situatie doet zich eveneens in zes gevallen voor.

Tabel 3.5: Gemiddelde cijfers ($\bar{X}$) en t-waarden voor het aantonen van eventuele volgorde-effecten bij het oplossen van PMP's. (I = als eerste probleem aangeboden; II = als tweede probleem aangeboden)

studie-jaar	cijfer	PMP1(I) $\bar{X}$	PMP1(II) $\bar{X}$	t	PMP2(I) $\bar{X}$	PMP2(II) $\bar{X}$	t
3	produkt	6.27	6.17	0.25	4.73	5.97	-2.36*
	totaal	5.83	5.70	0.35	4.78	5.93	-2.91*
4	produkt	5.93	6.79	-1.26	6.84	7.05	-0.34
	totaal	5.96	6.28	-0.59	6.29	6.42	-0.28
5	produkt	6.05	6.04	0.01	6.54	6.08	0.76
	totaal	5.93	5.76	0.30	5.97	5.62	0.74

* $p<.05$

Alleen bij de derdejaars studenten zijn significante verschillen aangetroffen. De studenten die PMP2 (het moeilijke PMP) als eerste probleem kregen aangeboden hebben systematisch lagere produkt- en totaalcijfers gehaald dan de studenten die dit PMP als tweede probleem moesten oplossen.

Dezelfde analyses werden uitgevoerd over de produktcijfers behaald op de PPP's (tabel 3.6). Geen enkel verschil bleek statistisch significant, alhoewel er in vijf van de zes gevallen sprake was van hogere prestaties op het als tweede probleem aangeboden PPP.

Geconcludeerd kan worden dat, met uitzondering van de derdejaars studenten die PMP2 oplosten, de prestaties van studenten niet beïnvloed zijn door de volgorde waarin de problemen zijn aangeboden. Bij de analyses ten behoeve van de overige vraagstellingen hoeft daarom geen rekening te worden gehouden met de volgorde waarin PMP's en PPP's zijn opgelost.

Tabel 3.6: Gemiddelde cijfers ($\bar{X}$) en t-waarden voor het aantonen van eventuele volgorde-effecten bij het oplossen van PPP's. (I = als eerste probleem aangeboden; II = als tweede probleem aangeboden)

studie-jaar	cijfer	PPP1(I) $\bar{X}$	PPP1(II) $\bar{X}$	t	PPP2(I) $\bar{X}$	PPP2(II) $\bar{X}$	t
3	produkt	5.55	5.59	-0.22	4.83	5.00	-0.95
4	produkt	6.28	5.94	1.19	5.74	6.34	-1.78
5	produkt	5.83	6.07	-0.77	5.43	5.88	-1.25

3.4.2 Zijn de geïntroduceerde moeilijkheidsgraden voor PMP's en PPP's geldig? (vraagstelling 2)

De geldigheid van de moeilijkheidsgraden werd onderzocht door middel van een aantal multivariate variantie-analyses met herhaalde metingen. In totaal zijn drie van dergelijke analyses uitgevoerd. In tabel 3.7 worden de resultaten gepresenteerd van een analyse waarin de produktcijfers voor PMP1 en PMP2 de afhankelijke variabelen zijn.

Tabel 3.7: Multivariate variantie-analyse met herhaalde metingen op de dimensie "moeilijkheidsgraad". (Onafhankelijke variabelen: studiejaar en moeilijkheidsgraad; afhankelijke variabelen: produktcijfers op PMP1 en PMP2)

variatiebron	kwadratensom	df	variantie	F
Tussen subjecten				
studiejaar	19.99	2	10.00	3.79*
subj. binnen groepen	160.96	61	2.64	
Binnen subjecten				
moeilijkheidsgraad	0.00	1	0.00	0.00
studiej. x moeilijkh.	7.56	2	3.78	1.60
moeilijkh. x subj. binnen gr.	143.74	61	2.36	

* $p < .5$

De onderzochte steekproef werd gevormd door de subjecten die beide PMP's kregen aangeboden (groep 1 en 2 uit tabel 3.1).
Uit de tabel valt af te lezen dat er een significant verschil is tussen studiejaren (F = 3.79), terwijl een moeilijkheidsgraad-effect niet aanwezig is.

Tabel 3.8: Multivariate variantie-analyse met herhaalde metingen op de dimensie "moeilijkheidsgraad". (Onafhankelijke variabelen: studiejaar en moeilijkheidsgraad; afhankelijke variabelen: totaalcijfers op PMP1 en PMP2)

variatiebron	kwadratensom	df	variantie	F
Tussen subjecten				
studiejaar	13.38	2	6.69	3.87*
subj. binnen groepen	105.46	61	1.73	
Binnen subjecten				
moeilijkheidsgraad	0.23	1	0.23	0.15
studiej. x moeilijkh.	1.66	2	0.83	0.52
moeilijkh. x subj. binnen gr.	96.58	61	1.58	

* $p<.05$

In tabel 3.8 worden de resultaten gepresenteerd van een analyse waarin de totaalcijfers voor PMP1 en PMP2 de afhankelijke variabelen vormen. De onderzochte subjecten zijn dezelfde als in tabel 3.7. Evenals bij de voorgaande analyse is het studiejaar-effect statistisch significant (F = 3.87). Dit keer wordt ook een moeilijkheidsgraad-effect aangetroffen, hoewel niet significant (F = 0.15).

Uit tabel 3.7 en 3.8 kan geconcludeerd worden, dat het bij de PMP's aangebrachte onderscheid in moeilijkheidsgraad niet geleid heeft tot wezenlijk verschillende prestaties op deze management problemen.

Een gelijksoortige analyse werd uitgevoerd om de moeilijkheidsgraden van de PPP's op geldigheid te onderzoeken. De onderzochte steekproef bestond uit subjecten die beide PPP's hadden opgelost (groep 5 en 6 uit tabel 3.1). Tabel 3.9 laat een significant studiejaar-effect zien (F = 16.61), alsmede een significant moeilijkheidsgraad-effect (F = 14.43). Laatstgenoemd effect bevestigt de aanwezigheid van een verschil in moeilijkheidsgraad tussen beide PPP's. De gegevens in tabel 3.6 laten zien dat de prestatieverschillen tussen de PPP's in de goede richting liggen; in bijna alle gevallen wordt er hoger gepresteerd op PPP1 (eenvoudig probleem) dan op PPP2 (moeilijk probleem).

Tabel 3.9: Multivariate variantie-analyse met herhaalde metingen op de dimensie "moeilijkheidsgraad". (Onafhankelijke variabelen: studiejaar en moeilijkheidsgraad; afhankelijke variabelen: produktcijfers op PPP1 en PPP2)

variatiebron	kwadratensom	df	variantie	F
Tussen subjecten				
studiejaar	17.10	2	8.55	16.61*
subj. binnen groepen	32.94	64	0.51	
Binnen subjecten				
moeilijkheidsgraad	6.96	1	6.96	14.43*
studiej. x moeilijkh.	1.62	2	0.81	1.68
moeilijkh. x subj. binnen gr.	30.86	64	0.48	

* $p<.05$

3.4.3 Zijn er verschillen in oplossingskwaliteit tussen een PMP en PPP van dezelfde moeilijkheidsgraad? (vraagstelling 4)

Voor het beantwoorden van deze vraag werden de testprestaties van de subjecten uit groep 3 en 4 (tabel 3.1) geanalyseerd. In deze groepen zaten alle subjecten die zowel een PMP als een PPP van dezelfde moeilijkheidsgraad aangeboden hadden gekregen. Eventuele verschillen in prestatie tussen deze, inhoudelijk identieke, management problemen zouden toegeschreven kunnen worden aan de uiteenlopende wijze waarop met deze instrumenten getracht wordt probleemoplosvaardigheid vast te stellen. Omdat voor elk subject in de onderzochte steekproef beschikt werd over een cijfer op een PMP en een cijfer op een PPP, werden de testprestaties geanalyseerd met een multivariate variantie-analyse met herhaalde metingen op de dimensie "testmethode". De analyse werd afzonderlijk uitgevoerd voor groep 3 (studenten die de eenvoudige management problemen oplosten) en groep 4 (studenten die de moeilijke management problemen oplosten). Tabel 3.10 geeft de resultaten weer van de uitgevoerde analyse voor groep 3. Het testmethode-effect is niet statistisch significant ($F = 0.03$), hetgeen betekent dat de studenten uit groep 3 ongeveer dezelfde prestaties hebben geleverd op PPP1 en PMP1. Anders dan bij tabel 3.7 tot en met 3.9 is er geen sprake van een significant studiejaar-effect ($F = 1.00$). Tussen studiejaar en methode, daarentegen, valt een bijna significant interactie-effect te constateren ($F = 2.94$; $p = .07$). Blijkbaar is het methode-effect niet hetzelfde voor elk studiejaar.

Tabel 3.10: Multivariate variantie-analyse met herhaalde metingen op de dimensie "testmethode". (Onafhankelijke variabelen: studiejaar en testmethode; afhankelijke variabelen: produktcijfers op PMP1 en PPP1)

variatiebron	kwadratensom	df	variantie	F
Tussen subjecten				
studiejaar	4.37	2	2.18	1.00
subj. binnen groepen	74.57	34	2.19	
Binnen subjecten				
methode (PMP1/PPP1)	0.03	1	0.03	0.03
studiej. x methode	5.91	2	2.95	2.94
methode x subj. binnen gr.	34.19	34	1.01	

Een gelijksoortige analyse werd uitgevoerd voor groep 4. Elk subject in deze groep had zowel PMP2 als PPP2 aangeboden gekregen. De resultaten van deze analyse staan in tabel 3.11. Evenals bij de vorige analyse is het methode-effect niet statistisch significant ($F = 0.59$). Ook het interactie-effect tussen studiejaar en methode is niet significant ($F = 0.83$), wat betekent dat in elk studiejaar de prestatieverschillen tussen PMP2 en PPP2 ongeveer van dezelfde omvang zijn.

Tabel 3.11: Multivariate variantie-analyse met herhaalde metingen op de dimensie "testmethode". (Onafhankelijke variabelen: studiejaar en testmethode; afhankelijke variabelen: produktcijfers op PMP2 en PPP2)

variatiebron	kwadratensom	df	variantie	F
Tussen subjecten				
studiejaar	41.69	2	20.84	8.93*
subj. binnen groepen	72.40	31	2.34	
Binnen subjecten				
methode (PMP2/PPP2)	0.82	1	0.82	0.59
studiej. x methode	2.35	2	1.17	0.83
methode x subj. binnen gr.	43.59	31	1.41	

* $p<.05$

Het studiejaar-effect is zeer significant, wat wil zeggen dat de gemiddelde prestaties per studiejaar systematisch van elkaar verschillen.
De in tabel 3.10 en 3.11 gepresenteerde variantie-analyses zijn gebaseerd op herhaalde metingen; voor elk subject in de analyse werd over twee waarnemingen beschikt. Voor de bestudering van vraagstelling 4 leverde dit het voordeel op, dat eventuele verschillen in oplossingskwaliteit tussen PMP's en PPP's niet veroorzaakt zouden kunnen zijn door verschillen in probleemoplosvaardigheid tussen oplossers. Wel kan het inhoudelijk identiek zijn van PMP's en PPP's met dezelfde moeilijkheidsgraad de verschillen in oplossingskwaliteit beïnvloed hebben. Vandaar dat vraagstelling 4 ook nog onderzocht werd door de prestaties van de subjecten uit groep 1, 2, 5 en 6 te analyseren. De subjecten in deze groepen hebben allemaal of twee PMP's of twee PPP's gemaakt; dat wil zeggen inhoudelijk verschillende problemen. Een tweevoudige variantie-analyse werd toegepast om verschillen in oplossingskwaliteit op te sporen tussen PMP's en PPP's voor de subjecten uit groep 1, 2, 5 en 6. Tabel 3.12 bevat de resultaten van deze analyse voor de "eenvoudige" management problemen.

Tabel 3.12: Tweevoudige variantie-analyse met "studiejaar" en "testmethode" (PMP1 - PPP1) als onafhankelijke variabelen en "produktcijfer" als afhankelijke variabele.

variatiebron	kwadratensom	df	variantie	F
studiejaar	7.98	2	3.99	2.82
methode (PMP1/PPP1)	6.59	1	6.59	4.66*
studiejaar x methode	1.38	2	0.69	0.49
residu	181.21	128	1.42	

* $p<.05$

Anders dan in tabel 3.10 en 3.11 is in tabel 3.12 sprake van een significant testmethode-effect ($F = 4.66$). De prestaties op een PMP verschillen dus significant van de prestaties op een PPP. Bestudering van tabel 3.5 en 3.6 leert dat over het algemeen beter gepresteerd wordt op PMP1 dan op PPP1. Het studiejaar-effect is bijna significant ($F = 2.82$; $p = .06$). Het kleine interactie-effect ($F = 0.49$) geeft aan dat de gesignaleerde prestatieverschillen tussen PMP's en PPP's ongeveer even groot zijn in elk studiejaar.

In tabel 3.13 zijn de resultaten gepresenteerd voor de "moeilijke" management problemen. Het testmethode-effect is significant ($F = 17.08$) en wijst op het bestaan van significante verschillen in oplossingskwaliteit tussen PMP2 en PPP2. Tabel 3.5 en 3.6 geven uitsluitsel over de richting van die verschillen; geconstateerd

kan worden dat de prestaties op PMP2 over het algemeen hoger zijn dan die op PPP2. Eveneens werd een significant studiejaar-effect ($F = 13.88$) gevonden. Het interactie-effect tussen studiejaar en testmethode was niet statistisch significant ($F = 0.49$).

Tabel 3.13: Tweevoudige variantie-analyse met "studiejaar" en "testmethode" (PMP2 - PPP2) als onafhankelijke variabelen en "produktcijfer" als afhankelijke variabele.

variatiebron	kwadratensom	df	variantie	F
studiejaar	41.65	2	20.83	13.88*
methode (PMP2/PPP2)	25.64	1	25.64	17.08*
studiejaar x methode	1.33	2	0.67	0.44
residu	196.64	131	1.50	

* $p<.05$

De naar aanleiding van de vierde vraagstelling uitgevoerde analyses leveren tegengestelde informatie op. Analyse van de data afkomstig uit groep 3 en 4 met een herhaald metingen design, resulteerde in een niet significant testmethode-effect. Een normale tweevoudige variantie-analyse, uitgevoerd over de data uit groep 1, 2, 5 en 6, daarentegen, liet wél significante testmethode-effecten zien. Een mogelijke verklaring voor deze uiteenlopende resultaten kan wellicht gevonden worden in het zogenaamde "carry-over effect"; een bij herhaalde metingen designs veel voorkomend verschijnsel. Van een carry-over effect is sprake als de prestatie op een bepaald niveau van de onafhankelijke variabele van invloed is op de prestatie op een ander niveau. In het concrete geval van vraagstelling 4: het oplossen van een PMP kan de prestatie op het direct daarna aangeboden (en inhoudelijk identieke) PPP positief beïnvloed hebben. Maar ook het omgekeerde is mogelijk. Vermoeidheid en/of verveeldheid als gevolg van het oplossen van een PMP (een voor de studenten onbekend type management probleem), heeft de prestatie op het direct daarna aangeboden (en inhoudelijk identieke) PPP negatief beïnvloed.

3.4.4 Zijn er verschillen tussen de derde-, vierde- en vijfdejaars studenten met betrekking tot de prestaties op PMP's of PPP's? (vraagstelling 3)

Dat deze vraagstelling later aan de orde komt dan vraagstelling 4 vindt zijn oorzaak in het feit dat de uitgevoerde analyses ten behoeve van de beantwoording van vraagstelling 2 en 4 tevens gegevens opleverden voor de beantwoording van de derde vraagstelling. In alle uitgevoerde analyses ten behoeve van vraagstelling 2

en 4 was "studiejaar" een onafhankelijke variabele met drie niveau's (derde, vierde en vijfde studiejaar). Bij de bespreking van tabel 3.7 tot en met 3.13 is ook steeds ingegaan op de studiejaar-effecten. In vijf van de zeven besproken analyses was er sprake van een significant studiejaar-effect en in één van de resterende analyses van een bijna significant effect ($p = .06$). De conclusie luidt, dat de verschillen in kennis en vaardigheden tussen de jaargroepen voor een deel terug gevonden kunnen worden in de prestaties op de management problemen. Maar dat voorzichtigheid geboden is moge blijken uit tabel 3.5 en 3.6; vergelijking tussen de studiejaren van de gemiddelde cijfers op PMP's en PPP's leidt namelijk tot de constatering, dat de vijfdejaars vaak slechter gepresteerd hebben dan de vierdejaars. In par. 3.5 zal dieper worden ingegaan op deze constatering.

3.4.5 Zijn er verschillen tussen de derde-, vierde- en vijfdejaars studenten met betrekking tot de wijze waarop PMP's van een bepaalde moeilijkheidsgraad worden aangepakt? (vraagstelling 5)

Per studiejaar is voor beide PMP's nagegaan op welke wijze de studenten deze patiëntproblemen hebben aangepakt. Ten behoeve van de overzichtelijkheid wordt alleen gekeken naar de behandelingsvolgorde die studenten hebben aangebracht in de door hen gesignaleerde deelproblemen. Concreter: hebben de studenten de indeling van het probleemoplossingsmodel (zie par. 2.2) gevolgd? Inventarisatie van de gemaakte keuzes in de brugsegmenten van secties (zie par. 2.4.2.2) maakte het mogelijk om gevolgde oplosroutes globaal te reconstrueren. Figuur 3.2, 3.3 en 3.4 illustreren de oplosroutes van respectievelijk derde-, vierde- en vijfdejaars studenten voor beide PMP's.
Voor evaluatie-doeleinden werden alle mogelijke oplosroutes toegewezen aan één van de volgende categorieën:
- ideale oplosroute;
- aanvaardbare oplosroute;
- onaanvaardbare oplosroute.

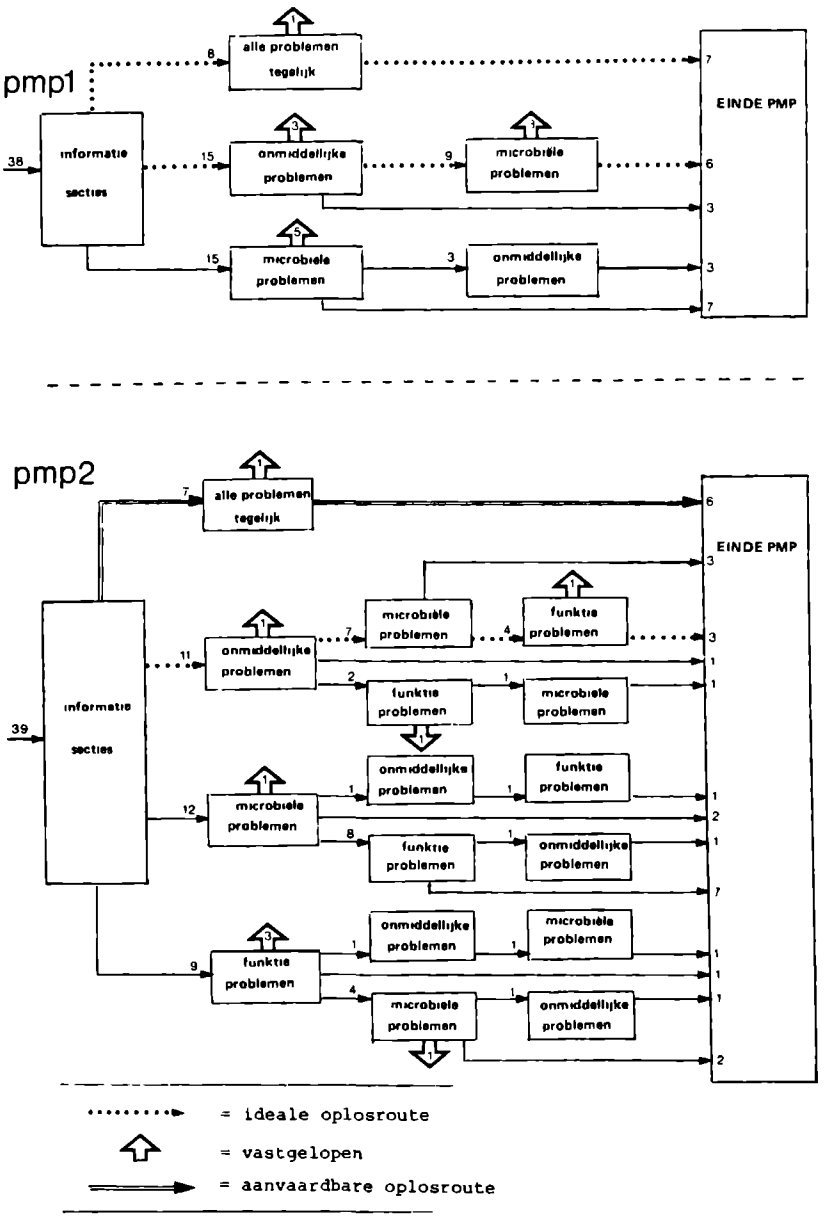
De ideale oplosroute was gevolgd als de behandelingsvolgorde van de gesignaleerde deelproblemen was gepland volgens de voorschriften van het probleemoplossingsmodel. Aanvaardbaar was een oplosroute als een student besloten had om de indeling van het probleemoplossingsmodel niet te volgen en alle problemen tegelijk aan te pakken. Tot de onaanvaardbare oplosroutes werden alle overige volgordes gerekend.

Om misverstanden te voorkomen dient hier opgemerkt te worden dat de kwaliteit van de oplosroute weliswaar van invloed is op de prestatie op het PMP, maar niet in die mate dat ideale oplosroutes een goede oplossing garanderen. Studenten die de ideale oplosroute hebben gevolgd zijn dus niet noodzakelijk de beste oplossers, wel de beste "aanpakkers". Een tweede belangrijke opmerking betreft een afwijking van de zo juist besproken indeling voor de kwaliteit van oplosroutes. Voor het eenvoudige PMP (PMP1) geldt dat er twee

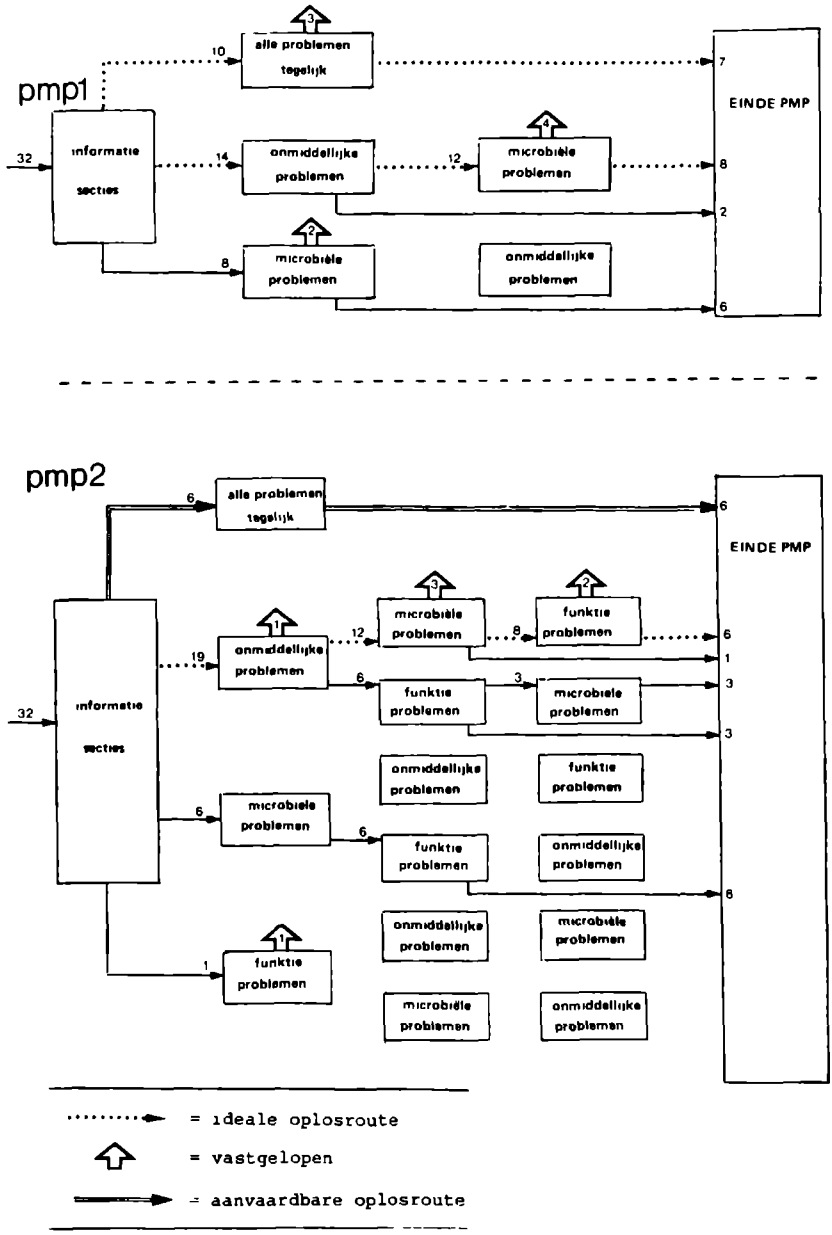
ideale oplosroutes zijn. De eerste ideale oplosroute is die welke voorgeschreven is door het probleemoplossingsmodel, de tweede is die waarin alle problemen tegelijk aangepakt worden. Alle andere oplosroutes, gevolgd bij het oplossen van dit PMP, zijn onaanvaardbaar. De categorie "aanvaardbare oplosroute" komt voor PMP1 dus te vervallen.

Uit onderstaande figuren kan eenvoudig afgelezen worden hoeveel studenten een oplosroute hebben gevolgd die als "ideaal" gekwalificeerd kan worden, als "aanvaardbaar" en als "onaanvaardbaar". Bijvoorbeeld, uit het schema van de oplosroutes van de derdejaars studenten in PMP1 (figuur 3.2) kan het volgende worden afgelezen:

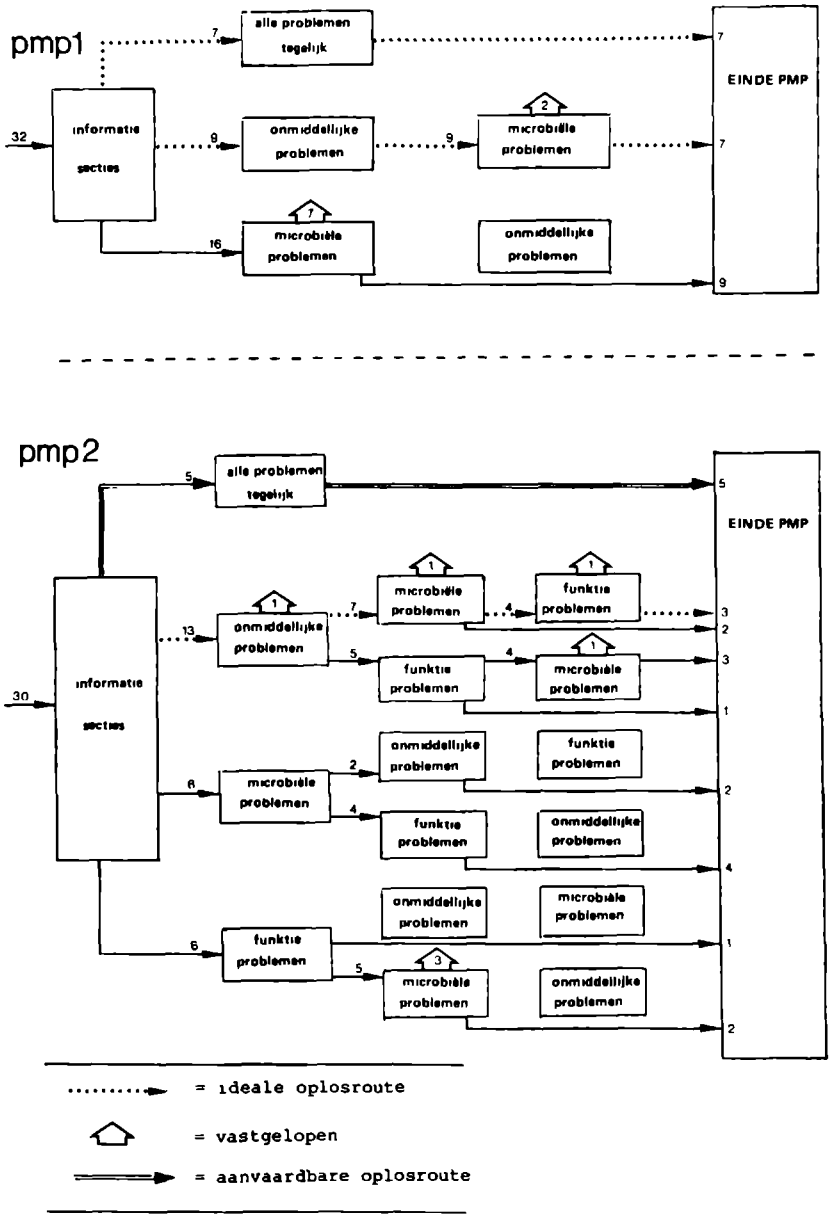
- achtendertig derdejaars zijn aan dit PMP begonnen;
- acht derdejaars besloten om na het inwinnen van informatie alle problemen tegelijk aan te pakken en één van hen liep daarbij vast;
- vijftien derdejaars losten na het inwinnen van informatie eerst de "onmiddellijke problemen" op en drie van hen raakten het spoor bijster in die fase. Negen van hen vervolgden het oplosproces met het aanpakken van de "microbiële problemen". Van deze negen liepen er weer drie vast, zodat er nog zes het probleem op reglementaire wijze beëindigden;
- vijftien derdejaars besloten om direct na het inwinnen van informatie over te gaan tot het oplossen van de microbiële problemen. In deze fase van het oplosproces beëindigden vijf van hen de simulatie op niet reglementaire wijze. Drie van de tien overgebleven studenten vervolgden het oplosproces met het aanpakken van de "onmiddellijke problemen". De zeven overige studenten beschouwden het probleem als beëindigd.



Figuur 3.2: Gevolgde oplosroutes van de derdejaars studenten op PMP1 en PMP2.



Figuur 3.3: Gevolgde oplosroutes van de vierdejaars studenten op PMP1 en PMP2.



Figuur 3.4: Gevolgde oplosroutes van de vijfdejaars studenten op PMP1 en PMP2.

De kwantitatieve gegevens uit de figuren kunnen overzichtelijker weergegeven worden in een frekwentietabel. Tabel 3.14 bevat de resultaten van deze inventarisatie voor PMP1 en tabel 3.15 voor PMP2.

Tabel 3.14: Kwalificatie van de gevolgde oplosroutes in PMP1. Tussen haakjes de relatieve aantallen (%) per studiejaar.

	studiejaar		
	3	4	5
ideale oplosroute	13 (34)	15 (47)	13 (34)
onaanvaardbare oplosroute	13 (34)	8 (25)	13 (34)
niet reglementair beëindigd	12 (32)	9 (28)	12 (32)
totaal aantal studenten	38	32	38

Tabel 3.14 laat zien dat even veel oplosroutes van derde- en vijfdejaars studenten gekwalificeerd konden worden als zijnde "ideaal" of "onaanvaardbaar". Ook het aantal niet reglementaire beëindigingen was in beide jaargroepen even groot. De oplosroutes van de vierdejaars studenten onderscheiden zich in positieve zin van die van de derde- en vijfdejaars. Bijna de helft (47%) van alle vierdejaars studenten volgde een oplosroute die als "ideaal" bestempeld kon worden. Onaanvaardbare oplosroutes werden door vierdejaars in mindere mate gevolgd dan door derde- of vijfdejaars studenten. Het percentage studenten dat vastliep in PMP1 was voor alle studiejaren ongeveer gelijk.

Tabel 3.15 laat ongeveer hetzelfde beeld zien als tabel 3.14. Het percentage vierdejaars studenten dat een ideale oplosroute heeft gevolgd is aanzienlijk groter dan dat van de derde- en vijfdejaars. Met betrekking tot het percentage aanvaardbare oplosroutes zijn de verschillen tussen de studiejaren kleiner, maar eveneens in het voordeel van de vierdejaars.
Het percentage onaanvaardbare oplosroutes was het kleinste voor de vierdejaars, terwijl tussen de studiejaren geen noemenswaardige verschillen optraden met betrekking tot het percentage niet reglementaire beëindigingen. Anders dan in tabel 3.14 is in tabel 3.15 een verschil zichtbaar tussen het derde en vijfde studiejaar; het percentage studenten uit laatstgenoemde studiejaar dat een ideale of aanvaardbare oplosroute volgde was iets groter dan dat van de derdejaars studenten.

Tabel 3.15: Kwalificatie van de gevolgde oplosroutes in PMP2. Tussen haakjes de relatieve aantallen (%) per studiejaar.

	studiejaar 3	4	5
ideale oplosroute	3 (8)	6 (19)	3 (10)
aanvaardbare oplosroute	6 (15)	6 (19)	5 (17)
onaanvaardbare oplosroute	21 (54)	13 (40)	15 (50)
niet reglementair beëindigd	9 (23)	7 (22)	7 (23)
totaal aantal studenten	38	32	38

De verwachting dat vijfdejaars studenten vaker goede oplosroutes zouden volgen dan vierde- en derdejaars en vierdejaars weer vaker dan derdejaars, kon slechts gedeeltelijk bevestigd worden door de gevolgde oplosroutes. Een duidelijke verklaring hiervoor ontbreekt. Wellicht dat in het onderwijs dat de vierdejaars genoten hebben meer de nadruk is gelegd op de wenselijkheid van gefaseerde behandelingsplanning dan in het onderwijs dat de vijfdejaars genoten hebben.

3.4.6 Is er sprake van een verband tussen prestaties op een PMP of PPP en prestaties op cognitieve toetsen? (vraagstelling 6)

Het vaststellen van de criteriumvaliditeit van de management problemen werd ernstig bemoeilijkt door het ontbreken van een geschikt criterium. De argumenten om "prestaties op cognitieve blokken" als criteria te gebruiken werden besproken in par. 3.2 onder punt c. Een selectieprocedure (zie par. 3.3.2.2) leverde tien cognitieve blokken op (tabel 3.4) die elk afzonderlijk als predictor zouden fungeren voor de prestaties op PMP's en PPP's. Per management probleem en per studiejaar werden Pearson correlatie coëfficiënten berekend tussen produkt-/procescijfers en cijfers voor de cognitieve blokken. Berekening per studiejaar was noodzakelijk, omdat in het derde en vierde studiejaar enkele cognitieve blokken nog niet waren aangeboden of nog niet waren afgerond met een toetsing. Tevens werden Pearson correlatie coëfficiënten berekend tussen de prestaties op elk cognitief blok en de prestaties op alle overige cognitieve blokken. Over deze correlaties werd een gemiddelde berekend. In de hieronder te presenteren tabellen zijn deze gemiddelde correlatie coëfficiënten per cognitief blok opgenomen in de kolom "GIC" (gemiddelde

intercorrelatie coëfficiënten). De GIC's kunnen de interpretatie van de correlaties tussen PMP's/PPP's en cognitieve blokken vereenvoudigen.

In tabel 3.16 tot en met 3.19 worden per studiejaar en per cognitief blok de besproken correlatie coëfficiënten gepresenteerd voor respectievelijk PMP1, PMP2, PPP1 en PPP2.

Tabel 3.16: Pearson correlatie coëfficiënten tussen testprestaties op cognitieve blokken en produkt-/procescijfers op PMP1. (PROD = produktcijfer; PROC = procescijfer; GIC = gemiddelde intercorrelatie coëfficiënt)

blok	3de-jaars (N=38)			4de-jaars (N=33)			5de-jaars (N=33)		
nr	PROD	PROC	GIC	PROD	PROC	GIC	PROD	PROC	GIC
100	-.02	-.09	.30	-.14	.14	.49*	-.08	-.13	.42*
103	-.07	-.13	.31	-.17	-.14	.48*	-.08	-.18	.45*
106	-.28	-.20	.22	-.08	.20	.27	.05	-.13	.38*
108	.05	.30	.11	-.32	.15	.36*	-.24	-.13	.26
203	.01	-.29	.31	-.03	.05	.57*	.02	-.24	.48*
206	.22	-.30	.15	-.18	-.13	.52*	.18	-.16	.39*
300				.02	.08	.54*	.16	-.12	.46*
306				.05	.32	.48*	-.06	-.17	.47*
312				-.08	.01	.53*	.15	.05	.25
403							.00	.07	.47*

* $p<.05$

De berekende correlaties tussen de produktcijfers op PMP1 en de cijfers voor cognitieve blokken (tabel 3.16) zijn geen van allen significant op het 5% niveau. Van de 25 berekende correlaties liggen er 21 tussen de waarden -0.20 en 0.20. Deze lage correlaties geven aan dat op grond van kennis over de prestaties op cognitieve blokken geen betrouwbare voorspellingen mogelijk zijn over het produktcijfer op PMP1. Met betrekking tot het procescijfer kan ongeveer hetzelfde geconstateerd worden. De negatieve correlaties zijn hier over het algemeen wat hoger dan bij de produktcijfers, maar te gering om er conclusies aan te verbinden. De cijfers op cognitieve blokken correleren onderling beter en zijn allemaal positief. Met uitzondering van de gemiddelde intercorrelatie coëfficiënten (GIC's) in het derde studiejaar zijn bijna alle GIC's significant op het 5% niveau.

Tabel 3.17: Pearson correlatie coëfficiënten tussen testprestaties op cognitieve blokken en produkt-/procescijfers op PMP2. (PROD = produktcijfer; PROC = procescijfer; GIC = gemiddelde intercorrelatie coëfficiënt)

blok	3de-jaars (N=39)			4de-jaars (N=33)			5de-jaars (N=31)		
nr	PROD	PROC	GIC	PROD	PROC	GIC	PROD	PROC	GIC
100	.22	-.16	.40*	-.15	-.04	.48*	.10	.01	.51*
103	.14	.22	.36*	.14	.06	.53*	.49*	.19	.46*
106	.06	-.11	.33*	.30	.51*	.17	.19	.00	.36*
108	.15	.03	.31	.07	-.10	.23	.12	.07	.28
203	.12	-.13	.42*	-.04	-.16	.53*	.26	-.06	.51*
206	.10	.11	.27	-.26	-.13	.41*	.38*	.16	.36*
300				-.04	.09	.53*	.20	-.15	.47*
306				-.16	.10	.45*	.13	-.15	.37*
312				-.30	-.37*	.51*	.09	-.26	.15
403							.26	-.05	.48*

* $p<.05$

De correlaties in tabel 3.17 wijken niet veel af van de in tabel 3.16 gepresenteerde correlaties. Met uitzondering van de correlaties tussen de cijfers voor blok 103 en 206, behaald door de vijfdejaars studenten, en het produktcijfer op PMP2, zijn de berekende correlatie coëfficiënten niet significant op het 5% niveau. De procescijfers blijken nog slechter voorspelbaar vanuit de prestaties op cognitieve blokken dan de produktcijfers. Met name geldt dit voor de prestaties van de vijfdejaars studenten. Slechts twee correlatie coëfficiënten tussen een cognitief blok en procescijfers zijn significant (blok 106 en 312). Correlaties tussen blokken onderling, daarentegen, zijn voor het merendeel wél significant.

Voor PPP's konden geen procescijfers berekend worden. Reden waarom in tabel 3.18 en 3.19 alleen correlatie coëfficiënten zijn opgenomen die berekend werden tussen blokcijfers en produktcijfers en tussen blokcijfers onderling.
Het beeld in tabel 3.18 en 3.19 wijkt nauwelijks af van dat in tabel 3.16 en 3.17. Het merendeel (80% en meer) van de berekende correlatie coëfficiënten tussen blokcijfers en produktcijfers is zeer laag ($-0.20 \leq r \leq 0.20$). Dit betekent dat vanuit de kennis van de prestaties op de cognitieve toetsen geen betrouwbare voorspellingen gedaan kunnen worden over de prestaties op PPP1 en PPP2. Anders gezegd: voor de vervaardigde PPP's kan geen criteriumvaliditeit vastgesteld worden.

Tabel 3.18: Pearson correlatie coëfficiënten tussen testprestaties op cognitieve blokken en produkt-cijfers op PPP1. (PROD = produktcijfer; GIC = gemiddelde intercorrelatie coëfficiënt)

blok nr	3de-jaars (N=40) PROD	GIC	4de-jaars (N=30) PROD	GIC	5de-jaars (N=34) PROD	GIC
100	.14	.38*	-.15	.46*	.12	.41*
103	.08	.43*	-.34	.50*	.07	.44*
106	-.10	.38*	.09	.51*	.12	.40*
108	-.13	.26	-.02	.54*	.12	.36*
203	-.01	.38*	-.11	.44*	.30	.50*
206	.05	.06	-.14	.51*	.46*	.32
300			-.16	.52*	.21	.36*
306			-.08	.47*	-.07	.40*
312			-.18	.46*	.08	.37*
403					.37*	.40*

* $p<.05$

Tabel 3.19: Pearson correlatie coëfficiënten tussen testprestaties op cognitieve blokken en produkt-cijfers op PPP2. (PROD = produktcijfer; GIC = gemiddelde intercorrelatie coëfficiënt)

blok nr	3de-jaars (N=40) PROD	GIC	4de-jaars (N=32) PROD	GIC	5de-jaars (N=31) PROD	GIC
100	-.21	.44*	-.08	.30	-.04	.55*
103	-.18	.46*	.03	.46*	.01	.48*
106	-.23	.44*	.06	.48*	-.35	.39*
108	-.15	.29	.12	.43*	-.19	.46*
203	-.14	.43*	.11	.37*	-.07	.55*
206	-.18	.15	-.17	.36*	.00	.30
300			.12	.48*	-.09	.39*
306			.03	.43*	-.05	.31
312			.08	.41*	-.08	.32
403					.10	.43*

* $p<.05$

Bijna alle gemiddelde intercorrelatie coëfficiënten in tabel 3.18 en 3.19 zijn statistisch significant op het 5% niveau, hetgeen betekent dat de prestaties op een cognitieve toets kennelijk beter geschikt zijn om prestaties op andere cognitieve toetsen te voorspellen dan prestaties op PPP's.

Ten aanzien van de zesde vraagstelling luidt de conclusie derhalve, dat noch voor de PMP's noch voor de PPP's criteriumvaliditeit kan worden vastgesteld. Op de mogelijke oorzaken hiervan wordt dieper ingegaan in de discussie (par. 3.5).

3.4.7 Wat zijn de meningen van de tandheelkunde studenten over het oplossen van PMP's en PPP's? (vraagstelling 7)

Studenten werden verzocht om direct na beëindiging van elk management probleem een aantal vragen te beantwoorden over hun ervaringen met en meningen over het opgeloste management probleem. Beantwoording van de vragen gebeurde door het aankruisen van een score op een vierpunts-schaal (bijlage 7).
Berekening van de gemiddelde score per vraag voor elk afzonderlijk management probleem, maakte het mogelijk om na te gaan of er sprake was van verschillen tussen de management problemen ten aanzien van de ondervraagde aspecten (tabel 3.20).

Tabel 3.20: Gemiddelde mening van de respondenten op de ondervraagde aspecten ten aanzien van PMP's en PPP's.

vraag	trefwoord	PMP1	PMP2	PPP1	PPP2
1	motivatie	2.24	2.26	2.40	2.56
2	kennis-verwerving	1.84	1.91	1.81	2.02
3	plezier	2.18	2.24	2.61	2.55
4	moeilijkheid	2.16	2.00	2.91	2.50
5	inzicht	2.17	2.43	3.32	2.99
6	betrokkenheid	2.40	2.22	2.55	2.54
7	gestruct. oplosproces	2.63	2.39	3.09	2.96
8	probleemoplosvaardigh.	2.29	2.27	2.73	2.85
9	simulatie	2.11	2.29	2.96	2.84

De meeste gemiddelden in tabel 3.20 liggen globaal tussen 2.00 en 3.00. De gemiddelde scores op PPP's zijn bijna altijd groter dan de gemiddelde scores op PMP's.
Vraag 1 gaat over de mate waarin management problemen de motivatie van de oplosser kunnen prikkelen. De respondenten waren van mening dat PPP's motiverender waren dan PMP's.

Geen van de management problemen werd positief beoordeeld op het aspect "kennisverwerving", zoals de gemiddelden voor vraag 2 laten zien.
Het plezier dat studenten beleefden aan het oplossen van een management probleem (vraag 3) was groter bij PPP's dan bij PMP's.
Het oplossen van PMP's werd moeilijker gevonden dan het oplossen van PPP's. PMP2 werd het moeilijkst bevonden en PPP1 het eenvoudigst (vraag 4).
De gemiddelde scores op vraag 5 laten zien dat studenten een beter inzicht hadden in de structuur van de PPP's dan van de PMP's. De structuur van PPP1 was voor de respondenten het meest duidelijk, terwijl het eenvoudige PMP (PMP1) het minst duidelijk werd bevonden.
Vraag 6 informeerde naar de betrokkenheid van de studenten bij de management problemen. Bij de PPP's bleek deze groter dan bij de PMP's.
Over het algemeen vonden de respondenten dat PPP's op meer gestructureerde wijze konden worden opgelost dan PMP's (vraag 7). PPP's werden ook hoger aangeslagen als instrument voor het meten van probleemoplosvaardigheid (vraag 8).
Ten slotte waren de respondenten van mening dat met PPP's het opstellen van een behandelingsplan beter nagebootst werd dan met PMP's (vraag 9).

Uit tabel 3.20 wordt niet duidelijk in hoeverre verschillen tussen management problemen echte verschillen zijn; dat wil zeggen geen toevallige verschillen. Eveneens wordt in die tabel geen informatie gegeven over mogelijke verschillen in opvatting tussen de respondenten uit de verschillende studiejaren. In tabel 3.21 worden de resultaten gepresenteerd van t-toetsen, uitgevoerd om verschillen tussen management problemen ten aanzien van de enquêtevragen statistisch te toetsen. Vergelijkingen tussen PPP's en PMP's van dezelfde moeilijkheidsgraad laten bijna altijd hogere gemiddelde scores zien voor de PPP's. In tabel 3.21 is dit zichtbaar in het positief zijn van de t-waarden. Behalve voor vraag 1, 2 en 6 zijn die t-waarden significant op het 5% niveau. Dit betekent dat de respondenten van mening zijn dat PPP's in vergelijking met PMP's eenvoudiger en plezieriger zijn om op te lossen. Daarnaast zouden ze een duidelijker beeld verschaffen van de problematiek en op meer gestructureerde wijze opgelost kunnen worden. Ten slotte zouden ze beter geschikt zijn om probleemoplosvaardigheid vast te stellen en een betere simulatie zijn van de werkelijkheid (het opstellen van een behandelingsplan voor een echte patiënt).
Vergelijkingen binnen een bepaalde testmethode (PMP1 vs PMP2 en PPP1 vs PPP2) leveren slechts twee significante t-waarden op. Studenten vonden PPP1 eenvoudiger op te lossen dan PPP2, alsmede dat PPP1 hun een duidelijker beeld verschafte van de aanwezige problematiek dan PPP2.
De conclusie luidt dat studenten een duidelijke voorkeur hebben voor het oplossen van PPP's, terwijl tussen PMP's en PPP's onderling geen respectievelijk bijna geen significante verschillen gevonden zijn.

Tabel 3.21: Toetsing (Student's t-test) van verschillen tussen management problemen met betrekking tot de enquête-vragen.

vraag (trefwoord)	PPP1-PMP1 t	df	PPP2-PMP2 t	df	PMP1-PMP2 t	df	PPP1-PPP2 t	df
1 motivatie	1.21	182	1.93	178	-0.13	174	-0.84	186
2 kennisverw.	-0.21	180	0.96	179	-0.63	175	-1.95	184
3 plezier	3.34*	182	2.32*	177	-0.41	174	0.48	185
4 moeilijkh.	5.85*	175	3.81*	176	1.20	167	3.69*	184
5 inzicht	4.78*	181	4.18*	177	1.56	172	2.81*	186
6 betrokkenh.	1.04	179	2.25*	175	1.27	169	0.03	185
7 gestruc.opl.	3.72*	178	4.34*	172	1.61	169	1.25	181
8 prob.oplosv.	3.29*	178	4.50*	174	0.09	169	-0.98	183
9 simulatie	6.15*	178	3.64*	173	-1.21	167	0.85	182

* $p<.05$

Verschillen tussen studiejaren ten aanzien van de beantwoording van de enquêtevragen werden eveneens getoetst met behulp van Student's t-test. In tabel 3.22 worden per enquêtevraag de t-waarden gepresenteerd die het resultaat zijn van vergelijkingen tussen derde- en vierdejaars studenten, vierde- en vijfdejaars studenten en derde- en vijfdejaars studenten.

Tabel 3.22: Verschillen tussen studiejaren met betrekking tot de beantwoording van de enquêtevragen. (Student's t-test)

vraag	trefwoord	3^e-4^ejaars t	df	4^e-5^ejaars t	df	3^e-5^e jaars t	df
1	motivatie	2.88*	242	1.17	214	4.19*	232
2	kennisverwerv.	1.15	243	2.03*	212	3.64*	231
3	plezier	0.16	243	0.97	213	1.19	232
4	moeilijkheid	-1.00	235	-0.42	208	-1.42	227
5	inzicht	-1.45	242	-0.45	212	-1.83	230
6	betrokkenheid	-0.67	239	0.69	207	0.19	230
7	gestruct.opl.	-1.89	237	1.63	206	0.06	225
8	probleemoplosv.	2.92*	236	0.69	207	3.49*	229
9	simulatie	2.20*	235	0.88	208	3.04*	227

* $p<.05$

Vergeleken met de vierde- en vijfdejaars studenten waren de derdejaars studenten significant positiever in hun oordeel ten aanzien van de motiverende werking van de management problemen, de geschiktheid ervan als instrument voor het meten van probleemoplosvaardigheid en de mate waarin ze de werkelijkheid nabootsen. Daarnaast bleek, dat derde- en vierdejaars studenten meer probleemoplosvaardigheid dachten te hebben verworven als gevolg van het oplossen van management problemen dan vijfdejaars studenten. Niet significant, maar wel in de verwachte richting, waren verschillen tussen de studiejaren met betrekking tot de moeilijkheid van de management problemen en het verkregen inzicht in de structuur van die problemen.

3.5 Discussie

In dit hoofdstuk is de aandacht vooral uitgegaan naar de validiteit van de vervaardigde management problemen. Drie soorten validiteit werden in beschouwing genomen, te weten: inhoudsvaliditeit, constructvaliditeit en criteriumvaliditeit. De inhoudsvaliditeit werd niet geëvalueerd. Verondersteld werd dat deze validiteit in voldoende mate aanwezig was in de problemen, omdat bij de constructie van de management problemen de doelstellingen van het studieprogramma centraal hadden gestaan (zie par. 2.4.2.3).

Een erg belangrijk aspect van management problemen is de moeilijkheidsgraad. Zonder kennis van de moeilijkheid van een probleem is het onmogelijk om vast te stellen of de oplosser ervan een aanvaardbaar niveau van probleemoplosvaardigheid bereikt heeft. De geldigheid van de gedefinieerde moeilijkheidsgraden voor de vervaardigde PMP's en PPP's werd onderzocht door na te gaan of de eenvoudige problemen inderdaad beter werden opgelost dan de moeilijke problemen.
Tabel 3.7 en 3.8 demonstreren dat de ingebrachte verschillen in moeilijkheid bij de PMP's niet geresulteerd hebben in significante verschillen tussen cijfers voor PMP1 en PMP2. Maar bij de PPP's zijn deze verschillen wél aangetroffen (tabel 3.9) en zijn ze bovendien in overeenstemming met de gedefinieerde moeilijkheidsgraden. Dat wil zeggen dat, over het algemeen, hogere cijfers werden behaald op PPP1 dan op PPP2 (tabel 3.6). In combinatie met het inhoudelijk identiek zijn van de PMP's en PPP's leidt de dubbelzinnigheid van deze resultaten tot de veronderstelling dat de testmethode zelf de kwaliteit van de oplossing beïnvloedt. Als de testmethode een PMP is, dan lijkt de moeilijkheid van ondergeschikt belang te zijn. Bij PPP's is de moeilijkheid van het probleem wél van invloed op de oplossingskwaliteit. Het is mogelijk dat de structuur van een PMP het oplosproces zodanig beïnvloedt, dat een bepaald prestatieniveau gemakkelijk bereikbaar is. Bijvoorbeeld als gevolg van het veelvuldig verstrekken van terugkoppeling aan oplossers en door het optreden van "cueing". Terugkoppeling dwingt oplossers om activiteiten te ontplooien die ze waarschijnlijk niet uit zichzelf zouden initiëren. Daarnaast worden verkeerde beslissingen al in een vroegtijdig stadium aan de

oplosser gemeld, waardoor de kans op een voldoende prestatie aanzienlijk wordt vergroot. Cueing doet zich voor als het nemen van een beslissing beïnvloed wordt door de verzameling van alternatieven waaruit de oplosser een keuze moet maken. De kans bestaat dat een oplosser daardoor betere beslissingen kan nemen dan wanneer elk alternatief eerst zelf bedacht moet worden. Wellicht is het voor PMP's nodig om de verschillen in complexiteit groter te maken als ze van verschillende moeilijkheidsgraad moeten zijn. In principe echter lijkt de methode voor het bepalen van de moeilijkheidsgraad (zie par. 2.4.2.3) bruikbaar te zijn.

Verschillen tussen studiejaren met betrekking tot de prestaties op PMP's en PPP's werden indicatief geacht voor constructvaliditeit. In de meeste analyses werden significante studiejaar-effecten aangetroffen (tabel 3.7, 3.8, 3.9, 3.11 en 3.13). Maar significante verschillen tussen studiejaren alleen, zijn niet voldoende om de aanwezigheid van constructvaliditeit te veronderstellen. Daarvoor is nodig dat de vierdejaars studenten hogere cijfers halen dan de derdejaars en dat de vijfdejaars weer beter presteren dan de vierdejaars studenten. Tabel 3.5 toont dat de vierdejaars studenten meestal een hoger gemiddeld produkt- of totaalcijfer behaalden op de PMP's dan de derdejaars studenten. De vijfdejaars studenten, daarentegen, presteerden meestal slechter dan de vierdejaars en afwisselend beter en slechter dan de derdejaars. Met betrekking tot de prestaties op de PPP's is de situatie iets gunstiger; behalve de vierdejaars presteren ook de vijfdejaars studenten beter dan de derdejaars. Maar evenals bij de PMP's zijn de gemiddelde cijfers van de vierdejaars studenten hoger dan die van de vijfdejaars (tabel 3.6). Voor dit verschijnsel kan een aantal mogelijke oorzaken genoemd worden:

1. De meest waarschijnlijke oorzaak lijkt te liggen in de uiteenlopende beheersingsgraad van de stof door de studenten uit de verschillende studiejaren. Het eenvoudige PMP/PPP is gebaseerd op tweedejaars doelstellingen en het moeilijke PMP/PPP op derdejaars doelstellingen. Het lijkt aannemelijk dat de afname van de management problemen, direct na de start van het studiejaar, er toe geleid heeft dat de derdejaars zich nog het beste de cognitieve kennis konden herinneren die nodig was voor het oplossen van het eenvoudige PMP/PPP. En even aannemelijk, dat de vierdejaars zich het beste de noodzakelijke stof voor het oplossen van het moeilijke PMP/PPP konden herinneren. Voor de vijfdejaars was het lang geleden dat de relevante stof bestudeerd was. Maar zij hadden het voordeel over de meeste ervaring te beschikken met betrekking tot het opstellen van behandelingsplannen.
2. Het is niet onmogelijk dat de vierdejaars studenten over meer of andere probleemoplosvaardigheid beschikken dan de vijfdejaars studenten.
3. Ten slotte kan de geringere motivatie van de vijfdejaars studenten (zie tabel 3.22) hun prestaties op de management problemen negatief beïnvloed hebben.

Vergelijking tussen studiejaren met betrekking tot de gevolgde oplosroutes was een andere manier om constructvaliditeit vast te stellen. De conclusies die uit tabel 3.14 (oplosroutes voor PMP1) en tabel 3.15 (oplosroutes voor PMP2) getrokken kunnen worden, zijn van dezelfde aard als de conclusies naar aanleiding van tabel 3.5 en 3.6. Zowel bij de oplossing van PMP1 als van PMP2 volgden de vierdejaars studenten betere oplosroutes dan de derde- en vijfdejaars studenten. Bij de oplossing van het eenvoudige PMP volgden derde- en vijfdejaars even vaak goede en slechte oplosroutes, maar bij de oplossing van het moeilijke PMP presteerden de vijfdejaars iets beter dan de derdejaars studenten.

De conclusie luidt dat enig bewijs is gevonden voor de aanwezigheid van constructvaliditeit in de geconstrueerde PMP's en PPP's.

De lage correlaties in tabel 3.16, 3.17, 3.18 en 3.19 geven aan dat het met de gebruikte, in par. 3.3.2.2 beschreven, methode niet gelukt is om criteriumvaliditeit aan te tonen. Mogelijke oorzaken voor de lage correlaties tussen prestaties op cognitieve toetsen en de management problemen zijn:

1. Het oplossen van management problemen vereist dat oplossers over nog andere vaardigheden beschikken dan dié welke nodig zijn voor het met succes afleggen van de geselecteerde cognitieve toetsen.
2. Bij de selectie van cognitieve toetsen zijn alleen toetstechnische criteria gehanteerd. Het vakinhoudelijke aspect van de cognitieve toetsen is buiten beschouwing gebleven. Achteraf bezien vormen de geselecteerde toetsen een tamelijk slechte representatie van de in het tandheelkundig curriculum aanwezige cognitieve vakken. In de selectie worden met name veel medisch-biologische vakken aangetroffen; vakken die voor het oplossen van tandheelkundige problemen wellicht van minder cruciaal belang zijn.
3. Het is mogelijk dat veel van de kennis, benodigd voor het met succes afleggen van de cognitieve toetsen, vergeten was.

Een aanwijzing voor eerstgenoemde mogelijkheid kan worden gevonden in de duidelijk hogere correlaties tussen cognitieve toetsen onderling. Prestaties op een cognitieve toets kunnen beter aangewend worden als predictor voor prestaties op andere cognitieve toetsen, dan als predictor voor prestaties op management problemen. Ook De Jong en Ferguson-Hessler (1982) komen tot een soortgelijke conclusie als ze beweren dat naast declaratieve kennis (kennis over feiten en principes op een bepaald vakgebied) en procedurele kennis (kennis over handelingen die in de omgang met declaratieve kennis zijn toegestaan) ook beschikt moet worden over selectiekennis en strategiekennis. Selectiekennis heeft volgens genoemde auteurs betrekking op het selecteren van voor de probleemoplossing relevante kennis uit het geheel aan kennis. Strategiekennis behelst kennis over oplosmethodes voor problemen. Inventarisatie van de doelstellingen van de cognitieve blokken (Studiegids Tandheelkunde, 1984-1985) liet zien dat het merendeel van de doelstellingen geformuleerd is in termen van "kennis van", "kunnen beschrijven van" en "kunnen aangeven van". Veel minder

vaak wordt een beroep gedaan op hogere cognitieve vaardigheden, zoals "inzicht", "analyseren" of "evalueren". Deze voor probleemoplossen belangrijke vaardigheden worden slechts zelden expliciet aangeleerd. In het tandheelkundig curriculum wordt alleen in het tweede studiejaar (blok 261: klinische tandheelkunde conserverend I) expliciet onderricht gegeven in het oplossen van tandheelkundige problemen. In een hoorcollege van één uur en in een aantal werkbesprekingen (maximaal vijf uur) worden de principes van de gebruikte heuristische methode (zie par. 2.2) uitgelegd en toegepast. Maar het is niet erg waarschijnlijk dat deze korte tijd voldoende is om van studenten vaardige probleemoplossers te maken. Het verdient daarom aanbeveling om de verworven kennis met betrekking tot het oplossen van problemen uit te bouwen, door deze toe te passen op vakspecifieke problemen binnen de cognitieve blokken. De kans dat studenten naast declaratieve en procedurele kennis ook nog selectie- en strategiekennis zullen verwerven, wordt daarmee aanzienlijk vergroot.
De belangrijkste conclusie die uit de correlatiestudie getrokken kan worden is dat geen criteriumvaliditeit vastgesteld kon worden, maar ook dat er aanwijzingen zijn dat de geselecteerde cognitieve toetsen geen geschikte predictoren waren. Voor het oplossen van management problemen is blijkbaar meer nodig dan declaratieve en procedurele kennis alleen.

Een andere belangrijke vraagstelling was of de vorm waarin het management probleem gestoken was van invloed is geweest op de kwaliteit van de oplossingen. Concreter geformuleerd: is er sprake van prestatieverschillen tussen PMP's en PPP's die inhoudelijk identiek zijn? Een vergelijking tussen de prestaties op een PMP en PPP van dezelfde moeilijkheidsgraad leverde geen significante verschillen op als een herhaald metingen design werd gebruikt (tabel 3.10 en 3.11). Omdat door het optreden van "carry-over effecten" (zie par. 3.4.3) de testprestaties wellicht beïnvloed zijn, werd dezelfde vraagstelling ook nog onderzocht door het uitvoeren van een normale tweevoudige variantie-analyse. Zowel voor de eenvoudige (tabel 3.12) als voor de moeilijke (tabel 3.13) management problemen werden significante verschillen gevonden tussen de prestaties op een PMP en een PPP. Vergelijking van tabel 3.5 met tabel 3.6 leidt tot de constatering dat de gemiddelde produktcijfers voor PMP's meestal hoger zijn dan die voor de PPP's. De al eerder in deze discussie ter sprake gekomen PMP-eigenschappen "terugkoppeling" en "cueing" lijken verantwoordelijk voor de hogere prestaties op PMP's. Het feit dat in PMP's veelvuldig terugkoppeling wordt verstrekt maakt ze ook bij uitstek geschikt als leermiddel (zie ook par. 1.6). De onmiddellijke terugkoppeling die na elke beslissing gegeven wordt, maakt dat het PMP een goede simulatie van de werkelijkheid is. Immers, het contact tussen tandarts en patiënt verloopt ook volgens een "actie-reactie patroon", waarbij de reactie een sturende invloed heeft op de volgende te ondernemen actie. Het optreden van "cueing" is een ongewenste eigenschap; in een reële situatie krijgt een tandarts geen lijst voor ogen waaruit de volgende te ondernemen actie gekozen kan worden. Cueing is een gevolg van het streven naar een

meer objectieve vaststelling van probleemoplosvaardigheid. Ten einde machinale beoordeling mogelijk te maken moeten de mogelijke oplosroutes vastgelegd zijn. Dit wordt bewerkstelligd door het aanbieden van keuzemogelijkheden in het PMP. Een nadeel hiervan is dat PMP's niet goed kunnen meten in hoeverre een oplosser in staat is om, op grond van reeds verzamelde informatie en beschikbare kennis en vaardigheden, zelf een volgende te ondernemen stap te kiezen. Dit creatieve aspect van probleemoplosvaardigheid kan beter vastgesteld worden met behulp van gecomputeriseerde PMP's (CPMP's), omdat cueing daar in mindere mate aanwezig is. In hoofdstuk IV zal uitgebreid worden ingegaan op deze CPMP's.

In tegenstelling tot de ervaringen van McGuire (1976) waren de studenten uit de onderhavige studie niet erg enthousiast over PMP's. Vergeleken met PPP's vonden ze PMP's minder plezierig om op te lossen, moeilijker, minder gestructureerd, minder geschikt om probleemoplosvaardigheid te meten en verder van de werkelijkheid staand. Hun mening dat PMP's minder gestructureerd zijn dan PPP's kan voortgekomen zijn uit de relatieve onbekendheid met PMP's en met het in de PMP's geïntegreerde probleemoplossingsmodel (zie par. 2.2). In het voordeel van de PPP-methode werkte het feit dat de studenten met deze methode vertrouwd waren. Daarnaast werden voor het opstellen van behandelingsplannen voor een PPP dezelfde formulieren gebruikt als die in het onderwijs gebruikt worden. Desondanks werden meestal hogere cijfers behaald op PMP's dan op PPP's. De vermoedelijke oorzaken daarvan kwamen in deze discussie reeds ter sprake. Duidelijk is geworden dat studenten van tevoren goed geïnstrueerd moeten worden over het werken met PMP's, vooral wanneer deze als toetsmiddel aangewend worden.

3.6 Conclusies en aanbevelingen

- Enig bewijs is gevonden dat de vervaardigde PMP's en PPP's constructvaliditeit bezitten.

- Prestaties op de geselecteerde cognitieve toetsen zijn geen goede voorspellers gebleken voor prestaties op management problemen. Op deze manier gedefinieerd kan geen criteriumvaliditeit worden toegeschreven aan de vervaardigde management problemen.

- Management problemen lijken een beroep te doen op specifieke vaardigheden. Dit onderstreept het belang van het geven van specifiek onderwijs op dit gebied.

- Het aangebrachte onderscheid in moeilijkheidsgraad is geldig gebleken voor de PPP's. De toegepaste methode voor het bepalen van de moeilijkheidsgraad van management problemen lijkt goed bruikbaar.

- Prestaties op PMP's waren over het algemeen hoger dan op inhoudelijk identieke PPP's. Vermoedelijk wordt dit veroorzaakt

door de "sturende" aspecten (terugkoppeling en cueing) die in PMP's aanwezig zijn.

- In tegenstelling tot wat verwacht werd, waren de studenten niet enthousiast over het oplossen van PMP's.

- De resultaten van deze studie moedigen de verdere ontwikkeling van tandheelkundige PMP's aan. Als PMP's gebruikt gaan worden als toets moeten studenten wel vertrouwd zijn met deze methode. Het regelmatig oefenen van het oplossen van PMP's is daarvoor noodzakelijk. Daarnaast verdient het aanbeveling om te onderzoeken in hoeverre PMP's efficiënte instrumenten zijn voor het verwerven van probleemoplosvaardigheid.

IV MICROCOMPUTER-SIMULATIE VAN TANDHEELKUNDIGE BEHANDELINGS-PLANNING

4.1 Inleiding

De relatief lage kostprijs in combinatie met de groter wordende werkgeheugens en het verschijnen van steeds meer interessante software, versnellen de opmars van de microcomputer in hoge mate. Ook in de Tandheelkunde zal binnen afzienbare tijd de microcomputer voor allerlei doeleinden worden ingeschakeld. In de tandartspraktijk zal dit voornamelijk de automatisering van de patiëntenadministratie betreffen. In het tandheelkundig onderwijs kan de microcomputer ingezet worden voor het verwerken van toetsen, het opslaan van toetsvragen ("itembanking"), het trainen van beoordelaars (zie deel I van dit proefschrift), het administreren van studievorderingen, enz. Alle hier genoemde toepassingen konden en kunnen ook gerealiseerd worden op de zogenaamde mainframes (grote computersystemen voor algemeen gebruik). Maar de microcomputer betekent een vereenvoudiging van diverse toepassingen doordat hij eenvoudig verplaatsbaar is, gemakkelijk in de bediening (geen ingewikkelde communicatietaal nodig) en eenvoudig uitbreidbaar met allerlei randapparatuur (printer, kaartlezer, videorecorder, tapedeck en diaprojector). Deze eigenschappen maken hem bijzonder geschikt als onderwijsmiddel. Als de computer gebruikt wordt als direct middel tot onderwijs en datgene wat geleerd moet worden binnen het systeem zelf aanwezig is, spreekt men van computergestuurd onderwijs (Moonen en Gastkemper, 1983). Hoewel computergestuurd onderwijs reeds langer bestond, lijkt het er op, dat met de ontwikkeling van de microcomputer deze vorm van onderwijs weer meer belangstelling krijgt van leerkrachten. Computergestuurd onderwijs (computer assisted instruction) is een overkoepelende term voor verschillende toepassingen van de computer als onderwijsmiddel. Moonen en Gastkemper (1983) noemen onder andere:

- drill and practice: gericht op het aanleren van (vooral) feitenkennis;
- tutorial: gericht op het verwerven van kennis en inzicht;
- dialogue and inquiry: gericht op ontdekkend leren;
- simulation and gaming: gericht op inzicht in, of (cognitieve) vaardigheid betreffende processen uit de werkelijkheid;
- problem solving: gericht op het oplossen van (cognitieve aspecten van) problemen.

Voor Tandheelkunde stammen de eerste toepassingen van computergestuurd onderwijs uit het begin van de jaren zeventig. De nadruk lag bij de meeste programma's op "simulatie" en "probleemoplossen". Sokolow en Solberg (1971), bijvoorbeeld, ontwikkelden een computergestuurd onderwijspakket dat studenten in staat stelde om hun diagnostische vaardigheden te oefenen cq te toetsen naar aanleiding van een door de computer gepresenteerd patiëntprobleem. Cassidy et al. (1972) ontwikkelden een programma waarmee

studenten hun bekwaamheid om tandpijn-problemen op te lossen konden vergroten en evalueren. Mullaney et al. (1976) ontwikkelden een computergestuurde onderwijsmethode, die de behandeling simuleerde van endodontische problemen.
Na 1977 kwam er een vrij plotseling einde aan de experimenten met computergestuurd onderwijs. De computer verdween echter niet uit het onderwijs, maar werd daar ingezet voor andere taken. Bijvoorbeeld voor het toewijzen van patiënten aan studenten, voor het bijhouden van de vorderingen van studenten, voor het evalueren van klinische vaardigheden, voor het vervaardigen en verwerken van toetsen, enz. Met een overkoepelende term worden dergelijke activiteiten aangeduid met "computerbeheerd onderwijs" (computer managed instruction).

De verschuiving van computergestuurd naar computerbeheerd onderwijs is zichtbaar in het aantal artikelen dat verschenen is over beide onderwerpen in de "Journal of Dental Education": een internationaal georiënteerd tijdschrift over het tandheelkundig onderwijs. In tabel 4.1 wordt voor elke jaargang tussen 1971 en 1983 gespecificeerd hoeveel artikelen verschenen zijn over computergestuurd en computerbeheerd onderwijs.

Tabel 4.1: Aantal artikelen per jaargang over computergestuurd (CGO) en computerbeheerd onderwijs (CBO) in de "Journal of Dental Education" tussen 1971 en 1983.

jaar	CGO	CBO	jaar	CGO	CBO
1971	2	0	1978	0	1
1972	4	0	1979	0	3
1973	1	1	1980	0	1
1974	1	1	1981	0	2
1975	2	1	1982	1	2
1976	3	0	1983	1	1
1977	1	1			

Opvallend is dat er na 1977 vier jaar lang geen artikelen zijn verschenen over computergestuurd onderwijs in de Journal of Dental Education. Alle publicaties over de toepassing van computers in het onderwijs gingen toen over computerbeheerd onderwijs.
In het Nederlands Tijdschrift voor Tandheelkunde werden in dezelfde periode slechts zes artikelen gepubliceerd waarin de toepassing van de computer in het onderwijs besproken werd. In geen van deze artikelen betrof het computergestuurd onderwijs.

In het medisch onderwijs in de Verenigde Staten heeft zich een soortgelijke ontwikkeling voorgedaan. In een artikel over het

inschakelen van microcomputers ten behoeve van computergestuurd onderwijs, noemen Schwartz en Hanson (1982) twee oorzaken voor de verminderde belangstelling hiervoor van medische opleiders. In de eerste plaats waren de kosten voor het onderhouden van bestaande systemen en het ontwikkelen van nieuwe computerprogrammatuur erg hoog. In de tweede plaats lag in de beginjaren van het computergestuurd onderwijs de nadruk te veel op het "technisch kunnen" van de computer, ten koste van de doelstellingen van het onderwijs en de behoeften van studenten en opleiders.
Schwartz en Hanson (1982) beschouwen de opkomst van de microcomputer als een tweede kans voor medische opleiders om gebruik te maken van de voordelen die computergestuurd onderwijs te bieden heeft. De auteurs noemen de volgende argumenten voor het gebruik van microcomputers ten behoeve van computergestuurd onderwijs:

1. lage aanschafprijs;
2. geringe operationele kosten;
3. geringe omvang van het geheugen, waardoor opleiders en programmeurs gedwongen worden om zich te richten op de leerdoelen in plaats van op technische "hoogstandjes";
4. beschikbaarheid van diverse hogere programmeertalen (onder andere: BASIC, PASCAL, PILOT, LISP), waardoor het schrijven van programma's relatief eenvoudig is.

Maar volgens Schwartz en Hanson zullen deze voordelen pas resulteren in een toepassing op ruime schaal van computergestuurd onderwijs als meer aandacht besteed zal worden aan het verstrekken van informatie over de toepasbaarheid van microcomputers in het medisch onderwijs.
Ook in het tandheelkundig onderwijs is een opleving te verwachten met betrekking tot het computergestuurd onderwijs als gevolg van de opkomst van de microcomputer. Een van de eerste toepassingen van microcomputergestuurd onderwijs is het project van Pryor en Racey (1982). Deze auteurs ontwikkelden een microcomputer simulatie voor het leren omgaan met levenbedreigende situaties. Een evaluatie van het systeem liet zien dat studenten erg snel leerden om "problemen" op een gestructureerde wijze aan te pakken. Verder bleek dat het systeem relatief goedkoop was (de aanschafkosten waren ongeveer $ 3000 terwijl de maandelijkse exploitatiekosten $ 20 bedroegen) en betrouwbaar (geen storingen tijdens de simulaties).

In dit hoofdstuk zal de ontwikkeling worden besproken van een programmapakket voor het opstellen van tandheelkundige behandelingsplannen door middel van simulatie met behulp van een microcomputer. Het is ontwikkeld in het Instituut Conserverende Tandheelkunde voor Volwassenen van de Katholieke Universiteit in Nijmegen. In navolging van Taylor et al. (1976) wordt het systeem aangeduid met CPMP (Computerized Patient Management Problem). Het ontwikkelde CPMP is inhoudelijk gebaseerd op het in hoofdstuk II beschreven PMP "Hendrik" en, evenals dit PMP, geschikt voor zowel instructie- als toetsdoeleinden. In paragraaf 4.2 worden de kenmerken van CPMP's besproken en in paragraaf 4.3 de constructie van het tandheelkundige CPMP "Hendrik". Paragraaf 4.4 is gewijd aan een pilotstudy naar het functioneren van dit CPMP.

4.2 Kenmerken van CPMP's

De kenmerken van CPMP's (Computerized Patient Management Problems) zullen hier besproken worden door ze te vergelijken met die van gewone PMP's. Het meest in het oog springende verschil is dat het oplossen van een PMP op papier gebeurt en van een CPMP op een beeldscherm. Echter, uit de hieronder te bespreken vergelijkingen tussen beide simulatievormen moge blijken dat een CPMP niet slechts een electronische versie is van een PMP.

1. Als in PMP's gebruik wordt gemaakt van de "latent image printing technique" (zie par. 1.5.3.1) voor het verbergen van de responsen, dan moeten die responsen liefst zo weinig mogelijk tekst bevatten. Het met de speciale viltstift ontwikkelen van lange responsen wordt door oplossers als vervelend ervaren. De als gevolg daarvan toenemende slordigheid bij het ontwikkelen kan gemakkelijk leiden tot onbedoeld ontwikkelen van bij andere opties behorende responsen. De beperkte omvang van de respons impliceert dat die doorgaans een zeer directe reactie is op beslissingen, acties en vragen van de oplosser. Dit is niet altijd in overeenstemming met de werkelijkheid, waar patiënten soms oppervlakkig en onsamenhangend antwoord geven op gestelde vragen en waar onderzoek soms dubbelzinnige resultaten oplevert. Bij CPMP's speelt dit bezwaar in mindere mate. Nadat een oplosser zijn keuze voor een bepaalde activiteit heeft ingetypt verschijnt de bijbehorende respons onmiddellijk op het scherm. De constructeur van het CPMP hoeft zich daarom minder te bekommeren om de lengte van de respons en kan die zodanig formuleren, dat de respons "verpakt" is in vage termen. Het is dan aan de oplosser om de respons naar waarde te schatten en het oplosproces op basis daarvan te continueren.

2. Het bij PMP's veelvuldig voorkomende verschijnsel "cueing" (zie par. 1.5.2 en 3.5) doet zich bij CPMP's in mindere mate voor. Daar waar de potentiële zoekruimte (zie par. 1.4) niet al te groot is en het aantal keuzemogelijkheden dus beperkt, kan het bij een CPMP aan de oplosser zelf worden overgelaten om een volgende activiteit te ontplooien. Het programma controleert of de ingetypte activiteit toelaatbaar is op dat specifieke moment in het oplosproces en gaat, indien dit het geval is, verder op de door de oplosser aangegeven weg. Als het een ontoelaatbare actie betreft krijgt de oplosser de mededeling dat een andere activiteit gekozen moet worden.

3. Om te vermijden dat oplossers steun zouden ondervinden van de volgorde waarin secties in een PMP zijn opgenomen, worden deze veelal in een willekeurige volgorde gezet. Daarnaast worden dikwijls "dummy-secties" opgenomen. Dat zijn secties waarin de oplosser nooit terecht kan komen omdat er nergens in het PMP naar verwezen wordt. Ze zijn opgenomen om te voorkomen dat een oplosser uit de afwezigheid van bepaalde secties conclusies zou trekken, die hem dichter bij de oplossing brengen. Genoemde maatregelen zijn er de oorzaak van dat PMP's vaak uit zeer veel

secties bestaan en dat het oplossen daardoor gepaard gaat met veelvuldig doorbladeren van het boekwerk. Door de oplosser kan dit ervaren worden als een hinderlijke onderbreking van het oplosproces. CPMP's kennen dit nadeel uiteraard niet; onmiddellijk nadat de oplosser te kennen heeft gegeven wat zijn volgende activiteit zal zijn komt er een respons van het systeem (antwoord van de patiënt, bevinding uit verricht onderzoek, andere activiteit) en een nieuwe opdracht of vraag.
Anders dan bij een conventioneel PMP, waar een oplosser slechts door heen en weer te bladeren een overzicht kan krijgen van de reeds verzamelde informatie, kunnen oplossers van CPMP's op elk moment een overzicht (op het beeldscherm of op papier) krijgen van de reeds ingewonnen informatie.

4. Het simulatiemodel van een PMP kan meestal aangeduid worden met de term "statisch" terwijl dat van een CPMP vaak "dynamisch" zal zijn. De term "statisch simulatiemodel" wordt door Hoffer et al. (1975) gebruikt voor simulaties waarin de klinische status van de patiënt ongewijzigd blijft tijdens het oplosproces. Simulaties waarin de klinische status van de patiënt zich kan wijzigen tijdens het oplosproces (bijvoorbeeld als gevolg van het verstrijken van de tijd of van de activiteiten van de oplosser), worden door genoemde auteurs "dynamisch" genoemd. Dynamische simulatiemodellen zijn realistischer dan statische en daardoor waarschijnlijk ook motiverender voor de oplosser. Dit kan een positieve invloed hebben op de validiteit van het simulatiemodel, gebruikt als meetinstrument, daar gemotiveerde oplossers eerder geneigd zullen zijn al hun kennis en vaardigheden aan te wenden dan ongemotiveerde oplossers.
Het in hoofdstuk II beschreven PMP "Hendrik" is een voorbeeld van een overgangsmodel; het heeft zowel statische als dynamische eigenschappen. Het dynamische karakter wordt bepaald door de wijze waarop terugkoppeling wordt gegeven over de resultaten van de uitgevoerde behandeling: de oplosser krijgt informatie over de gewijzigde klinische status van de patiënt. Het statische karakter schuilt in het feit dat de probleemoplosser niet in de gelegenheid wordt gesteld om eventuele nieuw ontstane problemen aan te pakken. Het in de volgende paragraaf te bespreken CPMP is een dynamisch simulatiemodel. Problemen die geïntroduceerd worden door de behandeling van andere problemen, worden bijgeschreven in een zogenoemde "problemenlijst" en moeten vervolgens worden opgelost.

5. In par. 2.4.1 onder punt 1 werd uiteengezet waarom het meestal niet mogelijk is om oplosroutes van opgeloste PMP's met grote nauwkeurigheid te reconstrueren. CPMP's kennen dit nadeel niet. Doordat elke activiteit van de probleemoplosser wordt vastgelegd in het achtergrondgeheugen van het computersysteem, kunnen oplosroutes achteraf perfect gereconstrueerd worden.
Een bijkomend voordeel van computersimulaties is dat ook informatie verkregen kan worden over de hoeveelheid tijd die elke activiteit van de probleemoplosser in beslag heeft genomen. Informatie hierover is belangrijk omdat de snelheid

waarmee gewerkt wordt een wezenlijk onderdeel is van probleemoplosvaardigheid (zie par. 1.3).

6. Doordat de gevolgde oplosroutes bij CPMP's precies bekend zijn, kan de betrouwbaarheid van de scoring groter zijn dan bij PMP's. Ook is het mogelijk dat de score direct na beëindiging van het CPMP wordt bepaald en aan de oplosser wordt verstrekt. Dat is aanzienlijk eenvoudiger dan bij PMP's, waar de gemaakte keuzes van probleemoplossers eerst nog ingevoerd moeten worden in een computer.

In voorgaande bespreking van de meest kenmerkende verschilpunten tussen PMP's en CPMP's zijn alleen dié aspecten aan de orde gekomen, die het CPMP tot een beter instrument maken voor het vaststellen van probleemoplosvaardigheid dan het PMP. Een bespreking van de belangrijkste nadelen mag echter niet achterwege blijven.

1. Tenzij over een groot aantal terminals of microcomputers beschikt kan worden, waardoor alle kandidaten tegelijk aan hetzelfde probleem kunnen werken, is een CPMP minder geschikt om als toetsmiddel gebruikt te worden. De kans is namelijk vrij groot dat probleemoplossers relevante informatie aan elkaar doorgeven, als ze het probleem moeten oplossen op verschillende tijdstippen.

2. Het vervaardigen van computerprogramma's voor CPMP's is een inspannend en tijdrovend karwei. Een extra handicap is het gemis aan standaardisatie in de automatisering. Programma's die voor een bepaald systeem geschreven zijn "draaien" niet zonder meer op andere systemen. Deze beperkte bruikbaarheid maakt programmatuur kostbaar.

3. Hoewel bij een CPMP de scoring veel gemakkelijker verloopt dan bij een PMP, is het opstellen van scorings-regels juist ingewikkelder. Dat vindt zijn oorzaak in het feit dat in CPMP's meestal een groter aantal oplosroutes gevolgd kan worden dan in PMP's. En dat is onder andere weer een gevolg van het dynamische karakter van CPMP's.

4.3 De constructie van een CPMP

4.3.1 De structuur

Het in deze paragraaf te bespreken CPMP is inhoudelijk identiek aan het in hoofdstuk II besproken PMP "Hendrik". Structureel zijn er opvallende verschillen. In het CPMP "Hendrik" genieten probleemoplossers een grote mate van vrijheid met betrekking tot te ondernemen activiteiten. In het gelijknamige PMP is een probleemoplosser verplicht om, eenmaal in een informatie-inwinsectie beland, alle opties die hem aantrekkelijk voorkomen achtereenvolgens te kiezen. Immers, er zijn geen mogelijkheden om na het

verlaten van een bepaalde informatie-inwinsectie hier, op een later tijdstip, weer in terug te keren. In het CPMP is geen sprake van een indeling in secties en is een probleemoplosser dus geheel vrij in het bepalen van de volgorde waarin informatie wordt opgevraagd.
Zoals in par. 4.2 werd besproken, is de omvang van het CPMP-probleem afhankelijk van de verrichtingen van de probleemoplosser. Als door een bepaalde behandeling nieuwe problemen ontstaan, dan worden deze toegevoegd aan de problemenlijst. De probleemoplosser wordt in de gelegenheid gesteld om ook deze problemen nog op te lossen. De grote mate van vrijheid die een CPMP biedt, blijkt ook uit de mogelijkheid om gesignaleerde problemen of geselecteerde oplossingen weer te verwijderen uit de problemen- respectievelijk oplossingenlijst, als er nog niet behandeld is.
De in vergelijking met een PMP minder starre structuur van een CPMP, laat de probleemoplosser meer vrijheid om het probleem op zijn eigen specifieke manier op te lossen. Het lijkt aannemelijk, dat hij daardoor meer blootgeeft van zijn probleemoplosvaardigheid dan door het oplossen van een PMP, waarin lang niet zo veel keuzes worden toegestaan.

4.3.2 De hardware configuratie

De simulatie wordt uitgevoerd op een Exidy^{R} Sorcerer^{R} microcomputer met 55K-byte RAM geheugen. Voor het laden van de vereiste programma's en het wegschrijven van de oplosroutes die door de probleemoplossers gevolgd worden, wordt gebruik gemaakt van een diskette station dat plaats biedt aan twee diskettes (5,25 inch, single sided/double density, 296 K-byte). Een Epson MX-82^{R} matrixprinter wordt gebruikt om informatie of overzichten op papier af te drukken. De apparatuur is verrijdbaar opgesteld.

4.3.3 De software

De software bestaat uit een uitvoeringsprogramma en vier bestanden (patiënt-, informatie-, problemen- en oplossingenbestand). Het geheel is ontwikkeld in het Instituut Conserverende Tandheelkunde voor Volwassenen van de Katholieke Universiteit te Nijmegen. Achtereenvolgens zullen in deze subparagraaf alle componenten van het programmapakket besproken worden.

Het uitvoeringsprogramma

Een belangrijke component van het programmapakket is het uitvoeringsprogramma (Sanders, 1984) dat geschreven is in Microsoft BASIC. Het uitvoeringsprogramma maakt het mogelijk om, op basis van vastgelegde informatie in de diverse bestanden, zeer nauwkeurig het opstellen van een behandelingsplan na te bootsen. Een dergelijke realistische benadering is mogelijk doordat grote nadruk is gelegd op flexibiliteit. Zo kunnen de meeste activiteiten zonder probleem op elk gewenst moment in het oplosproces

worden uitgevoerd. Als een probleemoplosser besluit om vlak voor een behandeling nog een röntgenfoto te nemen, dan is dat toegestaan. Het is ook mogelijk om geselecteerde problemen en/of oplossingen weer "af te voeren". Verder kunnen behandelingsplannen zo vaak herzien worden als wenselijk is.
Het uitvoeringsprogramma biedt probleemoplossers de keuze uit een groot aantal mogelijke activiteiten. Tabel 4.2 geeft een overzicht daarvan.

Tabel 4.2: Overzicht van de activiteiten die probleemoplossers kunnen ontplooien bij het oplossen van het CPMP.

- opvragen van de introduktietekst
- opvragen van informatie
- signaleren van een probleem in de verstrekte informatie
- selecteren van problemen
- formuleren van de doelstelling(en) van de behandeling
- bepalen van de volgorde waarin (geselecteerde) problemen moeten worden opgelost
- kiezen van oplossingen voor (geselecteerde) problemen
- opstellen van een concept behandelingsplan
- begroten van het behandelingsplan (niet beschikbaar in de geteste versie)
- uitvoeren van het behandelingsplan
- opvragen van alle gekozen oplossingen
- opvragen van de problemenlijst
- opvragen van het volgende scherm
- opvragen van een overzicht van alle mogelijke activiteiten die de probleemoplosser kan ontplooien
- kommentaar leveren op (onderdelen van) het CPMP
- verwijderen van een probleem of gekozen oplossing
- afdrukken van de informatie van het scherm op de printer
- tijdelijk beëindigen van het CPMP
- beëindigen van het CPMP

Andere functies van het uitvoeringsprogramma zijn:

- Koppeling van de informatie uit de diverse bestanden

Het uitvoeringsprogramma zorgt ervoor dat de uit verschillende bestanden afkomstige informatie, op de juiste wijze geïntegreerd wordt. Bijvoorbeeld: als een probleemoplosser bepaalde informatie opvraagt en daarin een probleem herkent, dan zoekt het uitvoeringsprogramma in het problemenbestand naar de problemen (inclusief afleiders) die bij het gekozen item uit het informatiebestand horen. Het uitvoeringsprogramma registreert vervolgens welk probleem door de probleemoplosser geïdentificeerd is en zoekt, als de probleemoplosser daar opdracht voor geeft, in het

oplossingenbestand naar oplossingen (inclusief afleiders) voor dat probleem. Als voor een bepaalde oplossing gekozen is, zoekt het uitvoeringsprogramma in het problemenbestand naar de terugkoppeling die bij die keuze hoort en verstrekt die vervolgens aan de probleemoplosser. Het is ook mogelijk dat geen terugkoppeling verstrekt wordt; bijvoorbeeld als het CPMP voor toetsing gebruikt wordt. Het uitvoeringsprogramma "weet" of er terugkoppeling verstrekt moet worden door naar de specificaties in het patiëntbestand te "kijken".

- Vastleggen van de oplosroute

Elke activiteit die een probleemoplosser onderneemt wordt geregistreerd in een apart bestand, zodat het mogelijk is om het oplosproces achteraf exact te reproduceren.

- Bijhouden van de tijd

Het uitvoeringsprogramma registreert de tijdstippen waarop de simulatie begint en weer eindigt. Daarnaast wordt bijgehouden op welk tijdstip voor een bepaalde activiteit gekozen wordt. Deze informatie wordt in hetzelfde bestand opgeslagen als de keuzes voor activiteiten, waardoor het achteraf mogelijk is om na te gaan welke activiteiten de meeste tijd in beslag hebben genomen en welke de minste.

Het patiëntbestand

Behalve de introductietekst van het CPMP bevat het patiëntbestand onder andere informatie over het doel van het CPMP, over de wijze waarop behandelingsplannen begroot worden, over uitdrukkelijke wensen van de patiënt en over eventuele relaties tussen problemen.
Als het CPMP wordt afgenomen om de probleemoplosvaardigheid vast te stellen, dan moet er tijdens het oplosproces minder terugkoppeling verstrekt worden dan wanneer het CPMP voor instructiedoeleinden wordt gebruikt. In het patiëntbestand kan worden opgegeven met welk doel het CPMP wordt afgenomen. Het uitvoeringsprogramma houdt daar vervolgens rekening mee door, bijvoorbeeld, alleen strikt noodzakelijke terugkoppeling te verstrekken als het CPMP als toets wordt aangewend.
Het begroten van het behandelingsplan kan automatisch gebeuren of door de probleemoplosser zelf, al naar gelang de specificatie in het patiëntbestand. In het eerste geval komen kosten en tijd van de geplande behandelingen automatisch in het behandelingsplan te staan, in het andere geval moet de probleemoplosser zelf de kosten en tijd berekenen en invoeren. Niet-automatische begrotingen worden getoetst op aanvaardbaarheid voor de patiënt. Als de kosten een bepaald bedrag overschrijden, dan accepteert de patiënt het behandelingsplan niet en moet de probleemoplosser het plan wijzigen.
De uitdrukkelijke wensen van de patiënt hebben betrekking op verrichtingen die de patiënt per se niet of wel uitgevoerd wil

zien. Als met deze wensen geen rekening wordt gehouden accepteert de patiënt het behandelingsplan niet.
Eveneens in het patiëntbestand opgenomen zijn de coderingen van de oplossingen die beslist niet en beslist wel in combinatie dienen voor te komen.
Tenslotte bevat het patiëntbestand informatie over de gerelateerdheid van bepaalde problemen. Op basis van die informatie "weet" het uitvoeringsprogramma, bijvoorbeeld, dat één behandeling diverse problemen kan oplossen.

Het informatiebestand

Het informatiebestand bestaat uit een groot aantal items die betrekking hebben op het verzamelen van informatie over de patiënt en zijn gebit. Ieder item bestaat uit een optie (vraag aan de patiënt, onderzoek bij de patiënt) en een respons (antwoord van de patiënt, bevinding uit onderzoek, reactie van de patiënt). Voor elk item is in het informatiebestand aangegeven of er één of meer problemen schuilen in de respons. Als dit het geval is worden tevens de identificatienummers vermeld van de problemen die het uitvoeringsprogramma op het scherm moet zetten, als een probleemoplosser te kennen heeft gegeven dat hij een probleem herkent in de respons. Achter elk identificatienummer staat een cijfer dat aangeeft of de probleemoplosser terugkoppeling mag krijgen over zijn probleemkeuze. Achter dat cijfer staat de tekst van de terugkoppeling.

Het problemenbestand

Het problemenbestand bestaat uit een groot aantal tandheelkundige of tandheelkundig relevante problemen, die al dan niet (afleiders) aanwezig zijn in het CPMP. Echt aanwezige problemen zijn voorzien van identificatienummers, die verwijzen naar oplossingen die het uitvoeringsprogramma op het scherm moet zetten als een probleemoplosser te kennen heeft gegeven dat hij oplossingen wil kiezen voor de gevonden problemen. Achter elk identificatienummer staat een cijfer dat aangeeft of er terugkoppeling verstrekt moet worden over de juistheid van de gekozen oplossing en de tekst van de terugkoppeling.

Het oplossingenbestand

In het oplossingenbestand zijn alle oplossingen (acceptabele én onacceptabele) opgenomen voor de in het CPMP echt aanwezige problemen. Voor elke oplossing staat vermeld hoeveel tijd er mee gemoeid is en wat de kosten er van bedragen. Verder wordt aangegeven of het uitvoeren van de oplossing tot een nieuw probleem leidt en, zo ja, het identificatienummer van dat probleem.

In tabel 4.3 wordt, aan de hand van een voorbeeld, gedemonstreerd op welke wijze informatie-, problemen- en oplossingenbestand onderling verbonden zijn. De respons van informatie-item nummer 28 bevat een (tandheelkundig) relevant probleem. Als een oplosser in

Tabel 4.3: Relaties tussen informatie-, problemen- en oplossingenbestand.

```
INFORMATIE BESTAND

#28.....................................item nr.
 Last van bloedend tandvlees?...........optie
 Bij het poetsen altijd.................respons
 @4.....................................aantal problemen
 1:1[Dit probleem is niet aanwezig.].....probleem nr:terugkopp.
 2:1[Dit probleem is inderdaad aanwezig;
    het wordt bijgeschreven in de lijst.]
 3:1[Dit probleem is niet aanwezig.]
 4:1[Dit probleem is niet aanwezig.]

PROBLEMEN BESTAND

#1.....................................probleem nr.
 hypertrophisch tandvlees...............probleemtekst
 @0.....................................afwezig probleem
#2
 gingivitis
 @4.....................................aanwezig probleem, met
                                        vier oplossingen.
 1:1[Onacceptabele oplossing; probleem
    blijft bestaan.]....................oploss. nr.:terugkopp.
 2:1[onacceptabele oplossing]
 3:1[uitstekende oplossing]
 4:1[onacceptabele oplossing; probleem
    blijft bestaan]
#3
 hypotrophisch tandvlees
 @0
#4
 poetstrauma
 @0

OPLOSSINGEN BESTAND

#1.....................................oplossing nr.
 gebit polijsten........................tekst oplossing
 10,37..................................tijd,kosten
 @1.....................................oplossing introduceert
                                        een (1) nieuw probleem
 2......................................identificatie nr. van het
                                        geintroduceerde probleem
#2
 poetsinstruktie rolmethode
 10,29
 @0
#3
 poetsinstruktie Bassmethode
 10,29
 @0
#4
 instruktie interdental brushes
 5,0
 @1
 2
```

antwoord op deze repons een "P" (van probleemherkenning) intypt, reageert het uitvoeringsprogramma met het op het scherm zetten van vier problemen, afkomstig uit het problemenbestand. De probleemoplosser geeft vervolgens aan welk(e) proble(e)m(en) volgens hem aanwezig is (zijn). Als het CPMP is ingericht voor instructie, reageert het uitvoeringsprogramma hierop door terugkoppeling te geven (zie de teksten tussen de rechte haken in het informatiebestand). In dit geval is alleen probleem 2 (gingivitis) aanwezig. Alleen dit probleem kan bijgeschreven worden in de problemenlijst (lijst van geïdentificeerde problemen). Als de probleemoplosser besluit om oplossingen te kiezen voor de geïdentificeerde problemen, zorgt het uitvoeringsprogramma er voor dat er vier mogelijke oplossingen op het beeldscherm komen voor probleem 2. Deze oplossingen (nummer 1 tot en met 4) staan in het oplossingenbestand. De probleemoplosser krijgt terugkoppeling (als het CPMP voor instructie gebruikt wordt) over de juistheid van zijn keuze voor een bepaalde oplossing. De teksten van die terugkoppeling staan in het problemenbestand. Oplossing 3 blijkt de enige acceptabele oplossing te zijn. In het oplossingenbestand wordt naast informatie over de benodigde tijd en kosten ook nog aangegeven (door middel van @0), dat deze oplossing geen nieuw probleem introduceert.

4.4 Een pilotstudy naar het functioneren van het CPMP

4.4.1 Inleiding

De flexibele structuur van het CPMP (zie par. 4.3.1) maakte het voor de constructeurs onmogelijk om te overzien of "alles" naar behoren functioneerde. Besloten werd om het programmapakket op proefondervindelijke wijze te testen. Het doel van de hieronder te bespreken pilotstudy was dan ook niet het vaststellen van probleemoplosvaardigheid of het geven van instructie over behandelingsplanning met behulp van een CPMP; primair ging de interesse uit naar het functioneren van het CPMP en naar de opgedane ervaringen van de deelnemers. De gegevens uit de pilotstudy zijn gebruikt om eventuele fouten in de programmatuur op te sporen en te herstellen en om het CPMP naar vorm en inhoud te optimaliseren.

4.4.2 Materiaal en methoden

Negen tandartsen van het Instituut Conserverende Tandheelkunde voor Volwassenen van de Katholieke Universiteit te Nijmegen hebben het CPMP opgelost. Geen van hen had eerder gecomputeriseerde patiënt management problemen opgelost, zodat hun ervaringen een goede indicatie zouden kunnen zijn voor de gebruikersvriendelijkheid van het programmapakket. Voorafgaand aan de simulatie kreeg elke deelnemer informatie over het doel van de studie en over de activiteiten die hij zou kunnen ontplooien om het probleem op te lossen. Er was geen tijdlimiet.

4.4.3 Resultaten

4.4.3.1 Functioneren van het CPMP

Alle in het CPMP beschikbare functies (zie tabel 4.2) bleken naar bevrediging te werken. Het uitvoeringsprogramma schakelde snel en foutloos over van de ene naar de andere functie. Onder alle omstandigheden verstrekte het uitvoeringsprogramma informatie aan de oplosser over de wijze waarop de simulatie gecontinueerd kon worden. Bijvoorbeeld door middel van mededelingen als: "uw keuze", "typ letter .. in gevolgd door RUN/STOP", "verkeerde invoer" of "voorbarige activiteit, kies opnieuw". Niettemin konden toch enkele tekortkomingen geconstateerd worden. Zo bleek dat de behandelingsvolgorde niet in het achtergrondgeheugen (diskette) werd vastgelegd. Ten behoeve van de pilotstudy werd deze onvolkomenheid provisorisch opgelost door behandelingsplannen automatisch door de printer te laten afdrukken, zodra een probleemoplosser gekozen had voor de functie "uitvoeren van het behandelingsplan".
Een andere tekortkoming betrof een verkeerde afhandeling van de procedure "automatische statuswijziging van problemen". In het CPMP kunnen problemen de volgende status hebben:
- "herkend" (het probleem is herkend in de opgevraagde informatie);
- "geselecteerd" (de oplosser heeft het probleem geselecteerd voor de eerstvolgende behandelingsronde);
- "gepland" (het probleem is opgenomen in een concept behandelingsplan);
- "behandeld" (voor het probleem is een afdoende behandeling gekozen en uitgevoerd).

Doordat in het CPMP enkele problemen nauw aan elkaar gerelateerd zijn, kan de behandeling van één van die problemen tevens leiden tot oplossing van de overige problemen. Dit betekent dat de status van die problemen automatisch gewijzigd moet worden bij de uitvoering van de juiste behandeling voor het gerelateerde probleem. Hoewel hiermee bij de constructie van het CPMP rekening was gehouden door dergelijke relaties tussen problemen vast te leggen in het patiëntbestand (par. 4.3.3), kregen sommige problemen toch een verkeerde status toegekend. Na iedere afname van het CPMP werd de status van de problemen gecontroleerd en werden, zo nodig, wijzigingen aangebracht in het uitvoeringsprogramma en/of patiëntbestand. Door de vele verschillende relaties tussen de problemen en de oplossingen voor die problemen lukte het pas bij de laatste afname om de statuswijziging van problemen foutloos te laten verlopen.

4.4.3.2 Ervaringen van de oplossers

Na beëindiging van de simulatie werd, in een open gesprek, elke probleemoplosser gevraagd naar zijn ervaringen met het oplossen van het CPMP. De ervaringen die door de meeste probleemoplossers gedeeld werden zullen hieronder besproken worden.

Alle probleemoplossers waren van mening dat het CPMP zeer gebruikersvriendelijk is. Ze ontleenden dit aan het feit dat ze, zonder vooraf noemenswaardige instructie te hebben gekregen, snel met de werking van het programma vertrouwd waren geraakt. Alleen bij de start ondervonden de meesten wat moeilijkheden als gevolg van de onbekendheid met de werkwijze. Eigenschappen die volgens de probleemoplossers in hoge mate hebben bijgedragen aan de gebruikersvriendelijkheid zijn:

- de mogelijkheid om op ieder gewenst moment een activiteit te onderbreken, een andere te ontplooien en vervolgens de onderbroken activiteit weer te continueren. Een voorbeeld kan de mogelijkheden verduidelijken. Het is voor een probleemoplosser mogelijk om tijdens het inwinnen van informatie over te gaan tot het kiezen van behandelingen voor de reeds geïdentificeerde problemen, een behandelingsplan op te stellen voor die problemen en dit plan uit te voeren. Als hij vervolgens besluit om nieuwe informatie in te winnen, dan keert het programma terug naar dát item van de informatieverzameling waar de activiteit eerder onderbroken werd.
- de mogelijkheid om op ieder gewenst moment een overzicht (op het beeldscherm of op papier) op te vragen van de status van de problemen of van geselecteerde behandelingen;
- de mogelijkheid om geïdentificeerde problemen en/of geplande verrichtingen weer af te voeren;
- de voortdurende informatie van het programma over de wijze waarop de simulatie gecontinueerd kan worden.

Alle probleemoplossers waren van mening dat het CPMP een goede benadering is van de realiteit en mede daardoor boeiend om op te lossen. Desalniettemin werden verscheidene suggesties gedaan om het CPMP nog realistischer te maken. Bijvoorbeeld: het voorstel om "doorvragen" mogelijk te maken. In reactie op de respons van de patiënt zou een andere vraag gesteld moeten kunnen worden, die dieper ingaat op het betreffende probleem of dit op andere wijze benadert. De in het CPMP aanwezige problemen zouden op deze wijze beter verborgen kunnen worden.
Een andere suggestie betrof het opnemen van de activiteit "bespreken van het concept-behandelingsplan met de patiënt". Een behandelingsplan zou pas uitgevoerd mogen worden als de patiënt er mee akkoord is gegaan. Hoewel echt bespreken van een behandelingsplan vooralsnog onmogelijk is, ligt het testen van een concept-behandelingsplan binnen de mogelijkheden. Op bescheiden schaal is dit in de huidige versie van het CPMP al aanwezig. Zo krijgt een probleemoplosser bij het uitvoeren van een behandelingsplan de mededeling dat bepaalde problemen vergeten zijn, als deze niet in het plan zijn aangetroffen. Dergelijke informatie wordt alleen gegeven voor díe problemen, waarvan de patiënt te kennen heeft gegeven dat hij die graag behandeld wil zien.
Een activiteit die nauw verband houdt met de aanvaardbaarheid van het behandelingsplan voor de patiënt, is het begroten van de kosten. Impliciet of expliciet kan in de op te vragen informatie vermeld zijn welk bedrag de patiënt maximaal wil besteden. Het is vrij eenvoudig om, uitgaande van een gemiddeld tarief, na te gaan

of de kosten van de geplande behandeling het maximaal te besteden bedrag overschrijden. In dat geval zou het programma de uitvoering van het plan moeten blokkeren en zou het behandelingsplan gewijzigd moeten worden door de probleemoplosser.
Een laatste hier te bespreken suggestie voor het realistischer maken van het CPMP werd eveneens door bijna iedere probleemoplosser geopperd, namelijk het verstrekken van visuele informatie. Het verstrekken van visuele informatie kan zowel versluierend als verhelderend werken. Versluierend, omdat de interpretatie van de visuele informatie aan de oplosser wordt overgelaten, waardoor aanwezige problemen minder gemakkelijk herkend worden dan wanneer de probleemoplosser ingevulde statusformulieren voorgelegd krijgt. Verhelderend, omdat woorden alleen soms niet voldoende zijn om een aanwezig probleem exact te definiëren. Stel bijvoorbeeld dat de parodontale status een "overhangende restauratie in de 46M" vermeldt. Zonder verdere informatie is het voor een probleemoplosser moeilijk om te beoordelen of dit probleem het beste kan worden opgelost door de restauratie bij te werken of door een nieuwe restauratie te leggen. Additionele visuele informatie, bijvoorbeeld in de vorm van een röntgenfoto, kan in zo'n geval uitsluitsel geven over wat de beste behandeling is. Behalve aan röntgenfoto's kan voor het verstrekken van visuele informatie nog gedacht worden aan gebitsmodellen, foto's of dia's en videoopnames.

Gebleken is dat een aantal probleemoplossers moeite had met de wijze waarop in het CPMP behandelingsplannen moeten worden opgesteld. Het stap-voor-stap opbouwen van een behandelingsplan strookte volgens hen niet met de werkelijkheid, waarin al die stappen niet of niet bewust uitgevoerd worden. Ook ongebruikelijk volgens hen, was het opstellen van afzonderlijke behandelingsplannen voor bepaalde groepen problemen in een behandelingsronde. Deze ongebruikelijke aanpak veroorzaakte bij bijna elke probleemoplosser enige verwarring, hetgeen tot uiting kwam in het kiezen van irrelevante of onmogelijke activiteiten in de fase direct volgend op het "inwinnen van informatie" en "herkennen van problemen". Een ander gevolg van de onbekendheid met deze manier van behandelingsplanning was de verkeerde interpretatie van sommige activiteiten uit het activiteitenoverzicht (tabel 4.2). Zo dacht een bepaalde probleemoplosser dat de activiteit "selecteren" exclusief bedoeld was voor het selecteren van dié problemen die geen uitstel konden velen.

Minder cruciale maar toch zinvolle opmerkingen hadden betrekking op de onoverzichtelijkheid van het activiteitenlijstje (tabel 4.2) en op het soms onduidelijke onderscheid tussen het "stellen van vragen aan de patiënt" en "het verrichten van onderzoek bij de patiënt" in de informatie-items.

4.4.3.3 Beknopte analyse van de oplosprocessen

Hoewel de afnames van het CPMP vooral bedoeld waren om het functioneren ervan te testen en de ervaringen van onervaren gebruikers te vernemen, is het illustratief voor de analysemogelijkheden van het CPMP om de oplosprocessen van de deelnemers aan de hand van enkele aspecten te beschrijven. Er zullen echter geen conclusies getrokken (mogen) worden over de probleemoplosvaardigheid van de betrokken oplossers. De belangrijkste reden hiervoor is dat aan belangrijke voorwaarden, op grond waarvan vergelijkingen tussen prestaties gerechtvaardigd zijn, niet voldaan is. Zo is bijvoorbeeld niet voldaan aan de eis van gestandaardiseerde testcondities. Pas bij de laatste afname werkte het CPMP-programma foutloos; bij alle voorgaande afnames was sprake van ernstige of minder ernstige fouten/storingen. Verder bleek, met name direct na de start, de behoefte aan hulp zeer verschillend te zijn bij de deelnemers. Een andere belangrijke reden is dat het CPMP was ingericht als "instructiepatiënt", waardoor oplossers veelvuldig geïnformeerd werden over het al dan niet juist zijn van door hen ondernomen activiteiten. Waarschijnlijk zullen de oplossingen daardoor op bepaalde aspecten (identificeren van problemen; kiezen van oplossingen) te weinig van elkaar verschillen om zinvolle uitspraken over de probleemoplosvaardigheid van de oplossers mogelijk te maken.

Het analyseren van de oplosprocessen is tamelijk eenvoudig omdat het uitvoeringsprogramma elke activiteit, door een probleemoplosser ontplooid om het probleem op te lossen, heeft "weggeschreven" naar het achtergrondgeheugen (zie par. 4.3.3). De "oplosbestanden" kunnen onder een tekstverwerkingspakket geladen worden, waarna de ontplooide activiteiten, met behulp van een sorteerroutine, op allerlei wijzen geordend kunnen worden. Hieronder worden de oplosprocessen van de deelnemers aan de pilotstudy geanalyseerd op de aspecten:
- benodigde tijd;
- hoeveelheid opgevraagde informatie;
- behandelingsvolgorde.

Benodigde tijd

Bij elke activiteit die ondernomen wordt "vraagt" het uitvoeringsprogramma de tijd op bij het systeem en slaat die samen met de codering van de betreffende activiteit op in het achtergrondgeheugen. Achteraf is dus zeer eenvoudig na te gaan hoe lang een probleemoplosser over de simulatie gedaan heeft. Interessanter echter, is dat nagegaan kan worden hoeveel tijd besteed is aan bepaalde onderdelen van het oplosproces. Dit zou bijvoorbeeld informatie kunnen opleveren over sterke en zwakke kanten van de probleemoplosvaardigheid van een oplosser of groep oplossers. In tabel 4.4 worden de tijden gepresenteerd van de deelnemers aan de pilotstudy op de onderdelen "informatieverzameling en probleemidentificatie", "kiezen van oplossingen" en "opstellen van behandelingsplannen". Voor de volledigheid wordt onder de rubriek

"overige" de tijd opgenomen die besteed werd aan het lezen van de introductie en evaluatie, het behandelen, het opvragen en printen van overzichten, het formuleren van doelstellingen en het geven van (schriftelijk) kommentaar op (onderdelen van) het CPMP.

Tabel 4.4: Overzicht van de tijd (minuten), besteed aan enkele onderdelen van het oplosproces door de deelnemers aan de pilotstudy. ($\bar{X}$ en s.d. zijn afgerond op hele minuten)

	probleemoplossers										
onderdeel	1	2	3	4	5	6	7	8	9	$\bar{X}$	sd
informatieverzameling en probl. identif.	64	43	38	65	41	53	37	33	36	46	12
kiezen van oplossingen	12	20	13	21	16	24	13	16	20	17	4
opstellen van beh. plannen	21	35	20	24	33	21	26	24	29	26	5
overige activiteiten	12	5	9	22	9	7	14	11	12	11	5
totaal	109	103	80	132	99	105	90	84	97	100	15

Uit tabel 4.4 blijkt dat de meeste tijd besteed is aan het opvragen van informatie en het identificeren van tandheelkundige problemen in de opgevraagde informatie. Overigens is er geen verband herkenbaar tussen de hoeveelheid opgevraagde informatie en de tijd daaraan besteed, zoals uit vergelijking van tabel 4.4 met tabel 4.5 blijkt. De verschillen in tijdbesteding op dit onderdeel zijn zeer groot; de snelste oplosser is bijna twee keer zo snel als de langzaamste oplosser. Waarschijnlijk zijn deze grote verschillen veroorzaakt door de verschillen tussen de probleemoplossers met betrekking tot het noteren van relevante informatie. Dit vermoeden lijkt bevestigd te worden door het feit dat bij de onderdelen "kiezen van oplossingen" en "opstellen van behandelingsplannen"

dergelijke grote verschillen niet zijn aangetroffen. In deze fasen van het oplosproces zijn dan ook geen aantekeningen gemaakt door de probleemoplossers.
Gemiddeld hadden de deelnemers 100 minuten nodig om het CPMP op te lossen.

Hoeveelheid opgevraagde informatie

In het CPMP kunnen in totaal 87 informatie-items opgevraagd worden. Deze items kunnen ondergebracht worden in rubrieken die gevormd zijn op basis van de aard van de informatie of op basis van het soort onderzoek dat verricht moet worden voor het verwerven van die informatie. Per informatie-rubriek is geïnventariseerd hoeveel items daaruit zijn opgevraagd door elke deelnemer. In tabel 4.5 wordt het aantal opgevraagde items per rubriek gepresenteerd, alsmede het gemiddelde en de standaardafwijking. Uit deze tabel blijkt dat er grote verschillen bestaan tussen de deelnemers met betrekking tot het aantal opgevraagde informatie-items. De verschillen tussen probleemoplossers die veel en weinig informatie opvragen zijn vooral geconcentreerd in de rubrieken "patiëntgegevens", "algemene gezondheid", "voedingsanamnese" en "extra-oraal onderzoek". Over de waarde van de items uit deze rubrieken voor het oplossen van het CPMP bestaan kennelijk grote meningsverschillen tussen de deelnemers.

Tabel 4.5: Aantal opgevraagde informatie-items per rubriek door elke probleemoplosser (gemiddelden en standaardafwijkingen zijn afgerond). A = absoluut aantal opgevraagde items per probleemoplosser; R = relatief (percentage) aantal opgevraagde items per probleemoplosser.

rubriek		probleemoplossers 1	2	3	4	5	6	7	8	9	$\bar{X}$	sd
patiënt-gegevens (13 items)		6	9	8	11	9	5	5	9	9	8	2
financ. mogelijkheden (4 items)		2	1	3	1	2	2	2	2	3	2	1
algemene gezondheid (8 items)		2	6	8	7	6	7	3	7	8	6	2
tandheelk. gezondheid (12 items)		6	7	10	11	10	10	7	8	11	9	2
voedings anamnese (11 items)		5	1	8	10	4	8	2	7	8	6	3
poets anamnese (8 items)		2	4	8	8	8	8	6	7	7	6	2
extra-or. onderzoek (9 items)		4	6	7	5	6	9	6	4	4	6	2
intra-or. onderzoek (15 items)		11	15	13	14	14	15	14	15	15	14	1
röntgen onderzoek (7 items)		3	5	4	5	3	2	6	3	6	4	1
totaal (87 items)	A	41	54	69	72	62	66	51	62	71	61	10
	R	47	62	79	83	71	76	59	71	82	70	12

Behandelingsvolgorde

De opgestelde behandelingsplannen kunnen onderling vergeleken worden door voor elk aanwezig probleem in het CPMP na te gaan of de probleemoplossers het herkend hebben, in welke behandelingsronde ze gepland hebben om het op te lossen, alsmede de volgorde (ten opzichte van andere problemen in dezelfde behandelingsronde) waarin behandeling van het probleem gepland is. In tabel 4.6 wordt een overzicht gegeven van de behandelingsvolgordes. De tabel moet als volgt gelezen worden:
- probleem 1 tot en met 17 zijn de in het CPMP aanwezige problemen;
- probleem 18 tot en met 23 zijn problemen die de probleemoplosser kan introduceren door het uitvoeren van bepaalde behandelingen;
- een lege cel geeft aan dat het betreffende probleem niet herkend is door een bepaalde probleemoplosser;
- coderingen in de vorm van x(y) geven aan in welke behandelingsronde een bepaald probleem wordt aangepakt (x) en in welke volgorde ten opzichte van andere (y), in die behandelingsronde op te lossen problemen;
- A-coderingen verwijzen naar dié problemen die "automatisch" zijn opgelost als gevolg van het oplossen van een ander (gerelateerd) probleem. Het getal achter de "A" verwijst naar het probleem waarvan de oplossing tevens een oplossing was voor het met "A" gecodeerde probleem. Als codering A ook nog voorzien is van een asterisk (*), dan betekent dit, dat het probleem weliswaar is opgelost maar door de probleemoplosser niet is herkend;
- "nvt" staat voor "niet van toepassing" en heeft alleen betrekking op problemen die kunnen ontstaan als gevolg van het oplossen van de aanwezige problemen (probleem 1 tot en met 17). De codering "nvt" geeft aan dat de aanwezige problemen zodanig zijn opgelost dat daardoor geen nieuwe problemen (probleem 18 tot en met 23) zijn geïntroduceerd.

In tabel 4.6 kan het volgende geconstateerd worden:
- Bijna alle aanwezige problemen (probleem 1 tot en met 17) zijn door elke probleemoplosser herkend. Alleen de problemen "fluoride-tekort" en "afwezigheid 37 en 47" werden veelal niet herkend.
- Behalve respondent 3 kozen alle probleemoplossers voor een gefaseerde behandeling. Het aantal behandelingsronden loopt echter sterk uiteen: twee oplossers (1 en 6) kozen voor twee behandelingsronden; vier (2, 5, 7 en 8) voor drie behandelingsronden en twee (4 en 9) voor zes behandelingsronden.
- Tussen de probleemoplossers bestaat grote overeenstemming over de fase waarin de problemen "gingivitis", "kaakgewrichtspijn", "avitaliteit 22" en "dwangbeet 17-46 en 27-36" het beste behandeld kunnen worden. Opvallend is dat de problemen waarvoor grote overeenstemming bestaat over de fase waarin ze behandeld moeten worden, juist dié problemen zijn die men in de eerste behandelingsronde wil aanpakken. Kennelijk worden deze problemen als meer urgent beoordeeld en wordt het minder relevant gevonden in welke behandelingsronde de overige (als minder urgent ervaren) problemen worden ingedeeld.

Tabel 4.6: Overzicht van de behandelingsvolgordes.

	probleemoplossers								
problemen	1	2	3	4	5	6	7	8	9
1.gingivitis	1(1)		1(2)	2(1)	2(1)	1(1)	1(1)	1(1)	1(1)
2.Fluoride tekort	2(8)			5(1)	2(2)				1(2)
3.kaakgewr. pijn	A*17	A*16	A-16	1(2)	2(4)	A-16	A-16	A-16	2(1)
4.infrabony pockets	1(2)	1(1)	1(3)	3(1)	2(3)	1(2)	1(5)	1(2)	4(1)
5.rookaanslag	2(7)	3(3)	1(4)	5(2)		1(3)	2(4)		
6.afwezigh. 37 + 47	A*17	A*16	A*16	A-3	A*3	A*16	A-16	A*16	A-3
7.overh.res. 46M	1(5)	2(1)	1(7)	2(2)	2(10)	1(6)	1(4)	2(1)	4(3)
8.overh.res. 34D	2(5)	2(2)	1(8)	2(3)	2(11)	1(5)	1(3)	2(6)	4(2)
9.sec.car. 23D	2(2)	2(3)	1(9)	4(3)	2(7)	1(11)	2(2)	2(5)	3(1)
10.sec.car. 24 MOD	2(3)	2(4)	1(10)	4(1)	2(8)	1(9)	2(3)	2(4)	3(2)
11.sec.car. 27MODB	A-17	A-16	A-16	A-3	A-3	A-16	A-16	A-16	A-3
12.sec.car. 44DO	2(4)	2(5)	1(11)	4(2)	2(9)	1(10)	2(1)	2(2)	3(3)
13.avitaliteit 22	1(4)	1(3)	1(5)	1(1)	1(1)	1(7)	1(6)	1(3)	5(1)
14.avitaliteit 35	2(1)	1(4)		1(3)	2(6)	1(8)	1(7)	1(5)	5(2)
15.fractuur kroon 22	2(6)	3(2)	1(6)	1(4)	2(5)	1(12)	3(1)	3(1)	5(3)
16.dwangbeet	A*17	1(2)	1(1)	A-3	A*3	1(4)	1(2)	1(4)	A-3
17.uitgroei 17 + 27	1(3)	A-16	A-16	A-3	A-3	A-16	A-16	A-16	A-3
18.verzw.knobbels 35		3(1)		6(1)	3(2)	2(2)	3(2)	3(2)	6(1)
19.diasteem 34-36	nvt	nvt	nvt	nvt	nvt	nvt	nvt	nvt	nvt
20.diasteem 21-23	nvt	nvt	nvt	nvt	nvt	nvt	nvt	nvt	nvt
21.onvol.res. 46MOD	nvt	nvt	nvt	nvt	nvt	nvt	nvt	nvt	nvt
22.onvol.res. 34MOD	nvt	nvt	nvt	nvt	nvt	nvt	nvt	nvt	nvt
23.onvol.ret. gietst. 35	nvt	nvt	nvt	nvt	nvt	nvt	nvt	nvt	nvt

4.4.4 Conclusies en aanbevelingen.

De twee belangrijkste doelstellingen van de pilotstudy waren het controleren van de juiste werking van alle in het CPMP beschikbare functies en het vernemen van de ervaringen van hen die het CPMP hebben opgelost. Gelet op deze twee doelstellingen kunnen de volgende conclusies getrokken worden:

1. Hoewel alle probleemoplossers (de laatste uitgezonderd) hinder hebben ondervonden van fouten in het programma, bleken alle functies over het algemeen naar bevrediging te werken.

2. De deelnemers aan de pilotstudy waren enthousiast over de gebruikersvriendelijkheid van het programma en het realistische karakter van de simulatie.

3. Gebleken is dat de probleemoplossers verschillend dachten over de betekenis van diverse functies voor het opstellen van een behandelingsplan. Uitvoerige toelichting vooraf over de beschikbare functies in het CPMP, is beslist noodzakelijk als de resulterende behandelingsplannen met elkaar vergeleken moeten kunnen worden.

4. Alle probleemoplossers zijn van mening dat de simulatie nog realistischer kan worden door het verstrekken van visuele informatie.

De positieve ervaringen van de deelnemers aan de pilotstudy moedigen nieuw onderzoek op het gebied van CPMP's aan. Allereerst dient onderzocht te worden of studenten, voor wie dergelijke instrumenten uiteindelijk bedoeld zijn, eveneens enthousiast zijn over het oplossen van CPMP's. Het is niet onmogelijk dat onervaren studenten hier andere opvattingen over hebben dan de ervaren deelnemers aan de pilotstudy. In par. 4.4.3, bijvoorbeeld, werd geschreven dat, door een aantal deelnemers aan de pilotstudy, het stap-voor-stap opbouwen van een behandelingsplan als niet realistisch werd ervaren. Studenten, echter, zouden deze aanpak juist als zeer realistisch kunnen ervaren, omdat de overeenkomst met de in het onderwijs aangeleerde aanpak groot is. Verder dient onderzocht te worden of en, zo ja, hoe CPMP's op zinnige wijze in het onderwijs kunnen worden ingeschakeld. Een belangrijk obstakel hiervoor is de tot op heden beperkte beschikbaarheid van computerapparatuur. Een reële toepassingsmogelijkheid kan worden gevonden in het gebruik van CPMP's als geïndividualiseerd instructiemiddel. Op basis van vrijwilligheid zouden studenten CPMP's kunnen oplossen om zich zo te bekwamen in het opstellen van tandheelkundige behandelingsplannen. Daarnaast zouden CPMP's op indirecte wijze een bijdrage kunnen leveren aan het onderwijs door ze te gebruiken voor het verifiëren van prescriptieve modellen van behandelingsplanning en voor het testen van PMP's. Verder zouden ze een uitgangspunt kunnen zijn voor de ontwikkeling van expertsystemen. In hoofdstuk V wordt uitgebreider op deze drie gebruiksmogelijkheden ingegaan.

V ALGEMENE DISCUSSIE EN AANBEVELINGEN

Het feit dat tandartsen, evenals medici, rechters en andere beroepsbeoefenaars die zich bezighouden met het oplossen van slecht gedefinieerde problemen, vaak geen beschrijving kunnen geven van het redeneerproces dat zij volgen om een voorgelegd probleem op te lossen, is geen bewijs voor de veelgehoorde opvatting dat de onderliggende cognitieve processen eerder een "kunst" zouden zijn dan een "kunde". De toenemende belangstelling voor probleemgeoriënteerd onderwijs is een aanwijzing dat steeds meer opleiders van mening zijn dat "probleemoplosvaardigheid" niet alleen het resultaat behoeft te zijn van langdurige ervaring maar ook direct aangeleerd kan worden. Een voorwaarde is dan natuurlijk dat men enig idee heeft van de cognitieve processen die zich afspelen als mensen proberen om problemen op te lossen. Onderzoek daarnaar mondt veelal uit in "descriptieve modellen" van het oplosproces. Een goed voorbeeld daarvan op het terrein van het medisch-diagnostisch denken is het werk van Elstein et al. (1978). Deze onderzoekers ontdekten dat artsen al in een zeer vroeg stadium enkele (drie à vijf) hypotheses formuleren over wat er aan de hand kan zijn met de patiënt en dat ze vervolgens proberen om voldoende bevestiging te vinden voor één daarvan. Verder bleek dat bij het stellen van een diagnose artsen erg veel waarde hechten aan de aanwezigheid van symptomen die een bevestiging zijn voor een hypothese terwijl ze andere informatie verwaarlozen.
Descriptieve modellen vormen veelal een uitgangspunt voor "prescriptieve modellen"; dat zijn beschrijvingen van hoe men het beste te werk kan gaan om problemen van bepaalde aard op te lossen. Doorgaans betreft het een verzameling heuristische procedures die de probleemoplosser kunnen helpen om tot een (niet gegarandeerd juiste) oplossing te komen. In hoofdstuk II (deel II) van dit proefschrift werden twee prescriptieve modellen besproken voor gebruik in het tandheelkundig onderwijs:
- de gemodificeerde probleemoplossingscyclus;
- Verdonschot's probleemoplossingsmodel.

Beide modellen zijn geconstrueerd om te dienen als hulpmiddel voor het systematisch leren opstellen van tandheelkundige behandelingsplannen. Het probleem met heuristische procedures is dat ze geen aanwijzingen geven die gegarandeerd leiden tot de oplossing van het probleem. Probleemoplossers krijgen slechts aanwijzingen die de kans op het bereiken van de juiste oplossing vergroten. Meestal is er sprake van een groot aantal wegen dat naar de oplossing leidt. Maar, dat niet alle wegen die naar de oplossing leiden aanvaardbaar zijn, werd met enkele voorbeelden geïllustreerd in par. 1.3. Om na te kunnen gaan of iemand voldoende probleemoplosvaardigheid bezit is het noodzakelijk om, naast de oplossing zelf, ook de gevolgde oplosroute te kennen. Aan probleemoplossers vragen om bij het oplossen hardop te denken en deze denkprotocollen vervolgens te analyseren is een mogelijke maar onpraktische en arbeidsintensieve manier om oplosroutes (en dus probleemoplosvaardigheid) vast te stellen. Ook de door de Nij-

meegse Subfaculteit Tandheelkunde gebruikte Papieren Patiënt Problemen (PPP's) bleken ongeschikt te zijn voor dat doel (par. 2.3). Het Patiënt Management Probleem (PMP) was een aantrekkelijk alternatief voor de PPP's. Vooral omdat het de oplosroutes van studenten kon vastleggen, maar ook omdat het een betere benadering was van de werkelijkheid en objectief gescoord kon worden.
Uit een studie naar de validiteit van twee geconstrueerde PMP's (hoofdstuk III) kwam naar voren dat er aanwijzingen waren dat de PMP's constructvaliditeit bezaten. De aanwezigheid van criteriumvaliditeit kon met behulp van de geselecteerde predictoren (een selectie van tien cognitieve toetsen) niet aangetoond worden. Maar, uit de vrij hoge correlaties tussen de cognitieve toetsen onderling en de zeer lage correlaties tussen deze toetsen en de PMP's werd geconcludeerd dat met de PMP's waarschijnlijk iets anders wordt gemeten dan met de cognitieve toetsen. Vooralsnog lijkt het echter onverstandig om uitspraken te doen over probleemoplosvaardigheid op basis van PMP-prestaties alleen.
Een andere constatering betrof de betere prestaties op PMP's in vergelijking met inhoudelijk identieke PPP's. De oorzaken daarvan zijn waarschijnlijk het optreden van "cueing" en het verstrekken van terugkoppeling tijdens het oplosproces in de PMP's (par. 3.5). Laatstgenoemde eigenschap maakt PMP's bijzonder geschikt voor instructiedoeleinden.

Het gebruik van PMP's als leermiddel beperkt zich niet tot het reguliere onderwijs; ook voor nascholing (in het kader van postacademisch onderwijs) kunnen ze effectief zijn. Tot die conclusie kwamen ook Marquis et al. (1984), die PMP's gebruikten voor de nascholing van huisartsen met betrekking tot cardiovasculaire problemen (hoge bloeddruk en angina pectoris). De belangrijkste constateringen van de onderzoekers waren:
- de deelnemende huisartsen verwierven kennis op het terrein van de hart- en vaatziekten door het achtereenvolgens oplossen van drie PMP's;
- het verstrekken van corrigerende terugkoppeling bevorderde het leren;
- een belangrijk deel van de nieuw verworven kennis werd onmiddellijk toegepast in de praktijk.

Een vervelend aspect van het gebruik van PMP's is de hoeveelheid tijd die het construeren van dergelijke instrumenten in beslag neemt. Ondanks de beschikbaarheid van gedetailleerde handleidingen voor de constructie van PMP's (McGuire et al., 1976; Verdonschot, 1983) blijft het ontwikkelen een tijdrovend karwei. Hoewel het mogelijk is om de aanwijzingen in dergelijke handleidingen zó gedetailleerd te maken dat de constructie van PMP's volledig gestandaardiseerd verloopt, is het risico niet denkbeeldig dat PMP's daardoor te veel op elkaar gaan lijken. Naarmate oplossers meer van dergelijke PMP's hebben opgelost kan de motivatie om goede prestaties te leveren afnemen.
Bass et al. (1981) werkten aan een nieuwe constructie-techniek voor PMP's die een mogelijke oplossing zou kunnen betekenen voor genoemde problemen. Een goed uitgangspunt voor de constructie van

een PMP is volgens deze auteurs een gestructureerd interview, aan de hand waarvan experts worden ondervraagd naar hun activiteiten en gedachten na de beschrijving te hebben vernomen van een hypothetische patiënt. De interviews leveren voldoende relevante informatie op om aan de hand daarvan een PMP te construeren. Bass et al. (1981) lieten op deze wijze PMP's vervaardigen door medisch specialisten en door studenten. Na afname bij een aantal proefpersonen konden geen verschillen in prestaties worden aangetroffen tussen de PMP's die vervaardigd waren door medisch specialisten en de PMP's die door studenten gemaakt waren.
Een andere mogelijkheid om de constructie van PMP's te vereenvoudigen is het uitgaan van bestaande casussen. Als opleiders eenmaal gewend zijn om over interessante casussen extra veel informatie te verzamelen, dan kan in korte tijd een grote verzameling patiëntgegevens opgebouwd worden aan de hand waarvan PMP's geconstrueerd kunnen worden. Een bijkomend voordeel is dat, anders dan bij hypothetische patiënten, tevens beschikt kan worden over visuele informatie. Dit laatste zal de simulatie nog realistischer maken.

Als de computer gebruikt wordt als medium voor het presenteren van een PMP, dan wordt van "CPMP" (Computerized Patient Management Problem) gesproken. Over de voordelen van CPMP's ten opzichte van PMP's is voldoende gezegd in par. 4.2. Op deze plaats wordt nader ingegaan op enkele activiteiten die een logisch vervolg zouden kunnen zijn op de ontwikkeling van tandheelkundige CPMP's.

1. Tandheelkundige CPMP's bieden de mogelijkheid om te onderzoeken in hoeverre tandartsen of tandheelkunde-studenten zich laten leiden door prescriptieve modellen bij het opstellen van behandelingsplannen. PMP's hebben deze mogelijkheid niet, omdat het oplosproces bij het oplossen van een PMP niet exact wordt vastgelegd en ze minder flexibel zijn dan CPMP's (zie par. 4.3.1), waardoor oplossers minder mogelijkheden hebben om het probleem op hun eigen specifieke manier op te lossen. Als op grond van een groot aantal afnamen zou blijken dat ervaren tandartsen (experts) op bepaalde punten systematisch van het prescriptieve model afwijken, dan zou dat een aanwijzing kunnen zijn voor onvolkomenheden in het model. Een andere mogelijkheid is dat de effectiviteit van een cursus, waarin studenten leren omgaan met een prescriptief model, wordt onderzocht door de oplosroutes te bestuderen die cursisten na afloop van die cursus gevolgd hebben bij het oplossen van een bepaald probleem.

2. Vooralsnog zal meestal over onvoldoende computerfaciliteiten beschikt kunnen worden om een groot aantal studenten gelijktijdig te laten werken aan een tandheelkundig CPMP. In dergelijke gevallen is men aangewezen op PMP's. CPMP's kunnen echter zinnig worden ingeschakeld ten behoeve van de constructie van PMP's door laatstgenoemde instrumenten inhoudelijk te testen alvorens tot het drukken van een groot aantal PMP-boekjes over te gaan.

3. CPMP's vormen een goed uitgangspunt voor de ontwikkeling van zogenaamde "expertsystemen". Expertsystemen zijn computersystemen die taken uitvoeren waarvoor specialistische kennis (expertkennis) nodig is. Ze worden gebruikt in kennisdomeinen waar men met strikt logisch redeneren alleen niet tot een goede oplossing kan komen en waar de beschikbare kennis een sturende functie vervult in het probleemoplossingsproces. Expertsystemen ontlenen hun kracht aan uitgebreide, aan menselijke experts ontleende, kennis. Meestal is die kennis gecodeerd in de vorm van honderden "als-dan"-vuistregels (heuristieken). De regels beperken het zoekproces door de "aandacht" van het programma te richten op de meest waarschijnlijke oplossingen (Lenat, 1984).
   Een goed voorbeeld van een domein waarin expertkennis vereist wordt, is de medische diagnostiek. Het medisch diagnostisch proces bestaat uit drie stadia:
   1. inwinnen van informatie;
   2. interpreteren van verzamelde informatie;
   3. besluitvorming.

   In stadium 1 en 2 is inschakeling van een computer mogelijk; stadium 3 is het terrein van de arts. Hij beslist uiteindelijk wat er met een patiënt aan de hand is en wat de beste behandeling is. Uiteraard kan de computer wel suggesties aandragen voor behandelingen en prognoses als functie van de ernst van de ziekte, de leeftijdsgroep of het geslacht. Bij inschakeling van de computer in stadium 1 kan men denken aan de automatische analyse van signalen (bijvoorbeeld van EEG's) en chemische processen en aan geautomatiseerde vragenlijsten. In het tweede stadium volgen menselijke probleemoplossers geheel andere methodes dan computerprogramma's. Uit het werk van Elstein et al. (1978) is gebleken dat artsen al in een zeer vroeg stadium zoeken naar symptomen die een bevestiging zijn voor opgestelde hypothesen. Computerprogramma's maken gebruik van een "beslissingsboom" of van een statistisch model en gaan veel breedvoeriger te werk. Het probleem met dergelijke programma's is dat ze ver af staan van het medisch denken van de arts en deze geen verklaring kunnen geven in patho-fysiologische termen, welke overweging tot de diagnose heeft geleid (Hasman, 1983). Expertsystemen, daarentegen, kunnen het denkproces van de arts simuleren en geven ook een verklaring voor de bereikte conclusies. Het meest besproken medische expertsysteem is MYCIN, ontwikkeld door Shortliffe (1976) van de Stanford University (Verenigde Staten). MYCIN diagnostiseert de oorzaak van bacteriële infectieziekten en geeft een advies voor een behandelingsplan. Het programma voert een dialoog met de gebruiker over eigenschappen van de patiënt en over de resultaten van eventuele laboratoriumtests. Tijdens de dialoog kan de gebruiker vragen waarom een bepaalde vraag wordt gesteld of op welke gronden een conclusie is getrokken. Het kennisbestand van MYCIN wordt gevormd door zogenaamde "produktieregels", die de volgende vorm hebben: als dit symptoom wordt waargenomen en ....., dan is deze ziekte met een kans van .. procent de oorzaak.

Op dit moment zijn er nog geen tandheelkundige expertsystemen, maar dat ze er zullen komen is waarschijnlijk. Naarmate meer tandartsen overgaan tot automatisering van hun patiëntenadministratie wordt de markt voor expertsystemen groter. Een eerste aanzet tot de ontwikkeling van tandheelkundige expertsystemen zijn de programma's die diagnoses kunnen stellen. Hyman en Doblecki (1983), bijvoorbeeld, ontwikkelden een programma dat kan adviseren over de noodzaak van een endodontische behandeling. Het betreft echter geen expertsysteem omdat het programma geen verklaring kan geven voor de gestelde diagnose.
In het tandheelkundig onderwijs kunnen expertsystemen worden ingezet voor instructie- en toetsdoeleinden. Ze zijn flexibeler dan CPMP's; laatstgenoemde instrumenten zijn gebaseerd op één concreet patiëntprobleem terwijl expertsystemen elk patiëntprobleem in een bepaald domein kunnen verwerken. Als eenmaal over een expertsysteem beschikt wordt dan vervalt daarmee de noodzaak om voor elk afzonderlijk patiëntprobleem een CPMP te construeren. Daarnaast zijn expertsystemen instructiever dan CPMP's, omdat ze de student op verzoek de beweegredenen verstrekken over de gestelde diagnose en het voorgestelde behandelingsplan.

Gelet op hetgeen in het voorgaande besproken is verdient het aanbeveling om:

1. de bruikbaarheid van PMP's en CPMP's ten behoeve van het post academisch onderwijs te onderzoeken;

2. te onderzoeken in hoeverre het haalbaar is om studenten PMP's te laten vervaardigen;

3. PMP's en CPMP's te baseren op bestaande patiëntproblemen;

4. na te gaan of het mogelijk is om prescriptieve modellen van de tandheelkundige behandelingsplanning te valideren aan de hand van CPMP's;

5. CPMP's in te schakelen om de inhoudelijke kwaliteit van PMP's-in-ontwikkeling vast te stellen;

6. te onderzoeken in hoeverre expertsystemen een zinvolle bijdrage kunnen leveren aan het tandheelkundig onderwijs in het algemeen en het onderwijs in de behandelingsplanning in het bijzonder.

NABESCHOUWING

In dit proefschrift zijn resultaten van onderwijskundig onderzoek beschreven in twee onafhankelijke delen, die als gemeenschappelijke noemer hebben het bevorderen van de kwaliteit van het tandheelkundig onderwijs.
Het eerste deel beschrijft een poging tot kwaliteitsverbetering van beoordelingen van practicumwerkstukken door het ontwikkelen en uittesten van een beoordelingsinstrument en een geïndividualiseerd trainingsprogramma ten behoeve van docenten.
Met het in het tweede deel beschreven onderzoek werd beoogd een bijdrage te leveren aan het verhogen van de kwaliteit van het tandheelkundig onderwijs door het ontwikkelen en uittesten van nieuwe instrumenten voor het "meten" van probleemoplosvaardigheid.
Wordt nu de balans opgemaakt dan blijft de vraag staan in hoeverre bovengenoemde instrumenten inderdaad de kwaliteit van het tandheelkundig onderwijs bevorderen. Voor een belangrijk deel zal dit afhankelijk zijn van de belangstelling en vakbekwaamheid van de in het onderwijs werkzame docenten. In dit verband moet gewezen worden op belangrijke ontwikkelingen die zich sinds enige tijd voltrekken in het wetenschappelijk onderwijs. Hierop zal nu beknopt worden ingegaan.

Ondanks de recente economische opleving zullen de universiteiten en hogescholen de komende jaren geconfronteerd blijven worden met bezuinigingen en (voor sommige studierichtingen) met afnemende studenten-populaties. Ook voor de studie Tandheelkunde zijn de jaren van groei voorbij. Enerzijds draagt hiertoe bij de TVC-operatie* (de subfaculteit in Utrecht gaat dicht en de twee Amsterdamse subfaculteiten moeten samen één opleiding gaan verzorgen), anderzijds kan de snel toenemende werkeloosheid onder tandartsen voor de minister aanleiding zijn om het aantal opleidingsplaatsen per jaar nog verder in te krimpen. Ten einde het gevaar van inkrimping of sluiting zo veel mogelijk af te wenden zullen de overgebleven subfaculteiten in Amsterdam, Groningen en Nijmegen trachten zich in positieve zin te onderscheiden.
Hoewel een dergelijke concurrentiestrijd op zich niet nadelig hoeft te zijn dreigt hier toch een groot gevaar. Omdat de schijnwerpers nu eenmaal eerder gericht worden op (resultaten van) wetenschappelijk onderzoek, ligt het voor de hand dat vooral de onderzoeksactiviteiten extra gestimuleerd zullen worden. Deze nadruk op het tandheelkundig onderzoek kan ten koste gaan van de kwaliteit van het onderwijs.

* Taakverdeling en concentratie wetenschappelijk onderwijs.

De kwaliteit van het onderwijs is vooral afhankelijk van het vakmanschap van de docenten. Dat "vakmanschap" meer omvat dan inhoudelijke kennis wordt duidelijk als nader geëxpliciteerd wordt wat bedoeld wordt met "kwaliteit van het onderwijs". Deze laat zich niet afleiden uit de leerplannen maar moet blijken uit de "opbrengst". De Groot (1983) zegt het als volgt: "De kwaliteit van het onderwijs dient beoordeeld te worden door de "output" van het onderwijs te beoordelen en te vergelijken, nadat men heeft vastgelegd wat men wil bereiken. Zonder doelstellingen geen standaard voor evaluatie en geen definitie van kwaliteitsverschillen."
Of de kwaliteit van het onderwijs beoordeeld wordt is behalve een kwestie van "willen" ook een kwestie van "kunnen". De vraag is dus niet alleen of docenten bereid zijn om het onderwijs dat ze verzorgen kritisch te evalueren en zo nodig bij te stellen, maar ook of ze daartoe in staat zijn. Lang niet alle docenten trekken conclusies over de kwaliteit van het door hen verzorgde onderwijs, als blijkt dat grote aantallen studenten de doelstellingen niet bereikt hebben. Meestal worden de studenten verantwoordelijk gesteld voor het falen. Deze handelwijze vindt zijn oorzaak niet zo zeer in het ontbreken van goede wil bij de docent, maar veeleer in het ontbreken van voldoende onderwijskundige kennis. Dit mag geen verbazing wekken aangezien tandartsen, zonder dat ze enige didactische scholing genoten hebben, als docent in het tandheelkundig onderwijs werkzaam kunnen zijn.
Bevordering van de onderwijsdeskundigheid van docenten is noodzakelijk om te komen tot kwaliteitsverbetering van het onderwijs. Het voorstel van de stuurgroep "docent-training" van de Subfaculteit Tandheelkunde in Nijmegen, om nieuwe instructeurs te verplichten een cursus "onderwijskunde" te volgen, moet om die reden worden toegejuicht.
Een bijkomend voordeel van de bevordering van de onderwijsdeskundigheid van docenten is de te verwachten betere samenwerking tussen docenten en onderwijskundigen. Docenten zullen beter in staat zijn om met onderwijskundigen van gedachten te wisselen over de wijze waarop de kwaliteit van het onderwijs bevorderd kan worden. Onderzoeks- en ontwikkelingswerkzaamheden, zoals bijvoorbeeld die welke in dit proefschrift beschreven zijn, kunnen daardoor veel beter aansluiten bij de behoeften van de docent. De kans dat onderzoeksresultaten slechts voor kennisgeving worden aangenomen en ontwikkelde produkten niet in het onderwijs worden ingevoerd, wordt daarmee aanzienlijk verkleind.

SAMENVATTING

DEEL I: BEOORDELEN VAN PRACTICUMWERKSTUKKEN

Het centrale thema in hoofdstuk I betreft de belangrijke rol die terugkoppeling speelt in het verwervingsproces van motorische vaardigheden. Ook in het tandheelkundig onderwijs is terugkoppeling, in de vorm van "kennis van de resultaten", van groot belang in met name de eerste fase van de verwerving van motorische vaardigheden. "Kennis van de resultaten" impliceert dat vorderingen vastgesteld kunnen worden. Meestal worden daarvoor "work-sample tests" gebruikt. Een work-sample test is een replicatie van een werksituatie (of onderdeel daarvan) die metingen oplevert aan de hand waarvan vastgesteld kan worden of een bepaalde vaardigheid beheerst wordt.
Het uiteindelijke doel van het tandheelkundig onderwijs is het opleiden tot "tandheelkundige competentie": een ingewikkeld samenspel van cognitieve, affectieve en motorische vaardigheden. Aan elk van die vaardigheden is een kwaliteits- en een kwantiteitsaspect te onderscheiden. Geconcretiseerd voor de motorische vaardigheden behelst het kwaliteitsaspect de gevarieerdheid en de mate van uitnemendheid van de dienstverlening. Het kwantiteitsaspect heeft betrekking op de hoeveelheid oefening die vereist is om van een constante kwaliteit verzekerd te kunnen zijn. Als de vaststelling van de motorische vaardigheden van dien aard is dat de student betrouwbare en valide terugkoppeling krijgt over zijn prestaties, dan kan de hoeveelheid oefening beperkt blijven. Er blijft dan meer tijd over om aandacht te besteden aan andere taken, waardoor studenten zich breder kunnen oriënteren en bekwamen.

Hoofdstuk II begint met een bespreking van het probleem dat inherent is aan het gebruik van work-sample tests, namelijk subjectiviteit in de beoordeling. De Groot (1971) noemt vijf specifieke moeilijkheden die zich kunnen voordoen bij beoordelingstaken:

1. signifisch effect: de beoordeling wordt beïnvloed door de opvatting van de beoordelaar over de beoordelingstaak;
2. halo-effect: de beoordeling wordt beïnvloed door opvallend goede of slechte aspecten van het te beoordelen werkstuk;
3. sequentie-effect: de beoordeling wordt beïnvloed door de kwaliteit van voorafgaande werkstukbeoordelingen;
4. persoonlijke vergelijkings-effect: de beoordeling wordt beïnvloed door algemeen menselijke, maar ook persoonlijke, neigingen tot specifieke beoordelingsverdelingen;
5. contaminatie-effect in engere zin: de beoordeling wordt beïnvloed door het feit dat de beoordelingsprocedure bewust of onbewust voor andere doeleinden wordt gebruikt dan voor beoordeling van het werkstuk.

Dat de genoemde beoordelaarseffecten zich ook voordoen bij werkstukbeoordelingen in het preklinisch tandheelkundig onderwijs wordt aangetoond aan de hand van diverse binnen- en buitenlandse studies. Pogingen, in die studies ondernomen, om de beoordelings-

kwaliteit te verbeteren waren vooral gericht op de volgende aspecten:
- formuleren van heldere, ondubbelzinnige prestatiecriteria;
- construeren van deugdelijke scoringssystemen;
- trainen van beoordelaars.

Vervolgens wordt aandacht besteed aan enkele subfacultaire (Subfaculteit Tandheelkunde van de Katholieke Universiteit in Nijmegen) onderzoeken naar de beoordelingskwaliteit van motorische vaardigheden. Geconcludeerd wordt dat de betrouwbaarheid van werkstukbeoordelingen in het preklinisch onderwijs veel te wensen overlaat. Gepleit wordt voor het construeren van nieuwe beoordelingsinstrumenten.

Hoofdstuk III is geheel gewijd aan de ontwikkeling van een zogenaamd "beoordelingsprotocol". Dat is een verzamelnaam voor de omschrijving, de beoordelingsmethode en het scoringsvoorschrift van te onderscheiden aspecten aan tandheelkundige werkstukken. Voor elk beoordelingsaspect van een werkstuk van het type "klasse II-tweevlakspreparatie voor amalgaam" wordt in het beoordelingsprotocol omschreven:
- aan welke eisen het werkstuk moet voldoen;
- hoe vastgesteld moet worden of het werkstuk aan die eisen voldoet;
- hoe de observatie in een score uitgedrukt moet worden.

In de laatste paragraaf wordt beschreven op welke wijze het functioneren van het beoordelingsprotocol is getest in een pilotstudy. De kwalitatieve (kritiek van docent- en student-beoordelaars) en kwantitatieve (inter- en intra-beoordelaarsovereenstemmingen) gegevens uit deze studie gaven aanleiding tot wijziging van het beoordelingsprotocol op diverse punten.

Centraal in hoofdstuk IV staat de ontwikkeling van een geïndividualiseerd trainingsprogramma voor beoordelaars. Door beoordelaars terugkoppeling te verstrekken over hun beoordelingsprestaties wordt geprobeerd de kwaliteit van beoordelingen te verbeteren. Twee voorafgaande voorwaarden voor functionerende terugkoppeling worden besproken:
1. Er moet een zodanige situatie geschapen worden dat de potentiële informatie-ontvanger zich openstelt voor terugkoppeling.
2. De informatie moet op zodanige wijze worden aangeboden dat de ontvanger er iets mee kan doen.

Beargumenteerd wordt dat geïndividualiseerde trainingen waarschijnlijk beter aan deze voorwaarden voldoen dan de gebruikelijke groepstrainingen.
Aandacht wordt besteed aan twee belangrijke vereisten voor geïndividualiseerde trainingen:
- de beschikbaarheid van een werkstukkenbestand;
- de automatisering van het trainingsprogramma.

In de laatste paragraaf wordt de opzet besproken van een onderzoek dat antwoord moet geven op de vraag of de kwaliteit van preklinische werkstukbeoordelingen verbeterd kan worden door beoordelaars gebruik te laten maken van het beoordelingsprotocol en

door ze te trainen in het beoordelen van practicumwerkstukken.

In hoofdstuk V wordt aandacht besteed aan de psychometrische betekenis van het begrip "betrouwbaarheid" en aan verschillende schattingsprocedures daarvoor (onder andere "split-half", "parallelvorm"). Voor het bepalen van de betrouwbaarheid van beoordelingen zijn die procedures meestal niet geschikt. Gebruikelijk is om de beoordelingsbetrouwbaarheid te schatten door het berekenen van de associatie of overeenstemming binnen of tussen een aantal beoordelaars. Het grootste deel van het hoofdstuk is gewijd aan een bespreking van de volgende statistische maten:
- coëfficiënt Kappa: een overeenstemmingsmaat voor nominale data;
- intraklasse correlatie coëfficiënt: een associatiemaat voor interval data;
- index T: een overeenstemmingsmaat voor interval data.

In de laatste paragraaf wordt uiteengezet wat verstaan wordt onder de validiteit van een beoordeling en op welke wijze die in het onderhavige onderzoek zal worden vastgesteld.

Met de beantwoording van drie vraagstellingen worden in hoofdstuk VI de resultaten besproken van het onderzoek naar het functioneren van het beoordelingsprotocol en het trainingsprogramma.
Aan de hand van de eerste vraagstelling werd onderzocht of de betrouwbaarheid van werkstukbeoordelingen toenam als het beoordelingsprotocol werd gebruikt in plaats van de gebruikelijke kenmerkmethode. De Kappa coëfficiënten die berekend werden over de subkenmerkbeoordelingen (beoordelingsprotocol) waren gemiddeld 63 procent groter dan de Kappa coëfficiënten berekend over de kenmerkbeoordelingen. Maar evenals bij de beoordelingen aan de hand van de kenmerkmethode was er bij beoordelingen aan de hand van het beoordelingsprotocol sprake van grote verschillen tussen de werkstukken met betrekking tot de beoordelaarsovereenstemming. Hieruit werd geconcludeerd dat, ondanks het gebruik van een beoordelingsprotocol, er waarschijnlijk sprake was van een onderscheid tussen moeilijk (kwalitatief juist voldoende of onvoldoende) en eenvoudig (duidelijk voldoende of onvoldoende) te beoordelen werkstukken. Een andere belangrijke constatering was dat het beoordelingsaspect "afwerking van de preparatie" ook met het beoordelingsprotocol niet betrouwbaar beoordeeld kon worden.
Met de tweede vraagstelling werd onderzocht of werkstukbeoordelingen aan de hand van het beoordelingsprotocol meer valide waren dan aan de hand van de kenmerkmethode. Ook deze vraag kon bevestigend beantwoord worden: de overeenstemming met het referentieoordeel (de kwaliteits-standaard) was gemiddeld twee keer zo groot op subkenmerk- als op kenmerkniveau.
Met de derde vraagstelling werd onderzocht of beoordelaars beter gingen beoordelen als gevolg van het deelnemen aan speciale trainings-sessies. Tussen hoeveelheid training en beoordelaarsovereenstemming kon geen positief verband ontdekt worden. Een positief verband werd wél aangetroffen tussen hoeveelheid training en de benodigde tijd per subkenmerk-beoordeling.

Tot slot worden in hoofdstuk VII enkele aanbevelingen gedaan voor de implementatie van het beoordelingsprotocol in het preklinisch onderwijs en voor een experimenteel onderzoek naar het effect van training op de beoordelingskwaliteit.

DEEL II: METEN VAN PROBLEEMOPLOSVAARDIGHEID

In hoofdstuk I wordt het terrein afgebakend waarop het onderzoek, dat in het tweede deel van deze dissertatie beschreven wordt, zich heeft toegespitst. Nadat een definitie is gegeven van de begrippen "probleem" en "probleemgeoriënteerd onderwijs" wordt uiteengezet dat "probleemoplosvaardigheid" niet zonder meer afgeleid kan worden uit het feit dat voor een bepaald probleem een oplossing is gevonden. Marshall's (1983) definitie biedt betere aanknopingspunten voor het vaststellen van probleemoplosvaardigheid. Volgens hem kan probleemoplosvaardigheid worden afgemeten aan de mate waarin een persoon er in slaagt om een probleem in zo kort mogelijke tijd op te lossen met minimale kosten en ongemak. Probleemoplosvaardigheid is dus alleen bereikbaar als bij het oplossen van problemen efficiënt te werk wordt gegaan. Slecht gedefinieerde problemen kunnen efficiënt worden opgelost door gebruik te maken van passende heuristische procedures.
Vier algemene klassen van heuristische procedures worden besproken:
- trial-and-error;
- nabijheids-methoden;
- fractionerings-methoden;
- kennis-gebaseerde methoden.

Een belangrijk deel van het hoofdstuk wordt in beslag genomen door een bespreking van de "papieren simulatie" als methode voor het vaststellen van probleemoplosvaardigheid. Papieren simulaties zijn beschrijvingen van situaties waarmee de beoefenaar van een bepaald beroep in de praktijk geconfronteerd kan worden. Wie de simulatie "doorwerkt" wordt geacht te denken en te beslissen alsof het een reële situatie betreft. Consequenties van beslissingen worden (meestal) onmiddellijk kenbaar gemaakt aan degene die het gesimuleerde probleem probeert op te lossen.
Nadat enkele vormen van papieren simulatie besproken zijn wordt uitvoerig aandacht besteed aan één vorm: het patiënt management probleem (PMP). Aan de hand van een soort basis-structuur van een PMP wordt besproken waarin PMP-varianten van elkaar kunnen verschillen. Vervolgens wordt aandacht besteed aan enkele nadelen van PMP's. Het hoofdstuk wordt afgesloten met de motivering van de keuze voor het PMP als instructie- en toetsmiddel ten behoeve van een preklinische cursus "behandelingsplanning" in het tandheelkundig onderwijs.

Hoofdstuk II begint met een bespreking van twee methoden voor het systematisch leren opstellen van tandheelkundige behandelingsplannen. De eerste methode is gebaseerd op de zogenaamde "empirische cyclus" en wordt momenteel in het tandheelkundig onderwijs gebruikt.

De tweede methode is recent ontwikkeld door Verdonschot (1982) en onderscheidt zich van eerstgenoemde methode door het verstrekken van "denkregels" aan de student en door het feit dat expliciet gebruik wordt gemaakt van de heuristische procedure "fractionering".
Vervolgens worden de tekortkomingen besproken van het "papieren patiënt probleem" (PPP) als methode voor het vaststellen van probleemoplosvaardigheid bij tandheelkunde studenten. Beargumenteerd wordt dat patiënt management problemen hier beter voor geschikt zijn.
Daarna volgt een uiteenzetting over de constructie van twee tandheelkundige PMP's. Daarbij wordt veel aandacht besteed aan de integratie van Verdonschot's probleemoplossingsmodel in de structuur van de management problemen. Tenslotte wordt ingegaan op de keuze van een methode voor het bepalen van de moeilijkheidsgraad van PMP's.

Gelet op het in hoofdstuk II uitgesproken vermoeden dat PMP's beter geschikt zouden zijn voor het vaststellen van probleemoplosvaardigheid bij tandheelkunde studenten dan PPP's, wordt in hoofdstuk III de validiteit van de PMP's vergeleken met de validiteit van (inhoudelijk identieke) PPP's. Omdat dé validiteit niet bestaat is nader gespecificeerd welke typen validiteit onderzocht worden.
In de studie werden twee PMP's en twee PPP's onderzocht. PMP1 en PPP1 zijn gebaseerd op tweedejaars doelstellingen en bevatten een identiek patiëntprobleem. PMP2 en PPP2 zijn eveneens inhoudelijk identiek maar berusten op derdejaars doelstellingen. In een verplichte toets losten alle derde-, vierde- en vijfdejaars studenten in het studiejaar 1983-1984 achtereenvolgens twee management problemen op.
De aanwezigheid van constructvaliditeit werd onderzocht door de prestaties van de studenten uit de drie studiejaren met elkaar te vergelijken. Een aanwijzing voor de aanwezigheid van constructvaliditeit werd gevonden in de hogere cijfers die vierde- en vijfdejaars behaalden voor zowel PMP's als PPP's dan de derdejaars. Maar dit gegeven wordt enigszins afgezwakt door het feit dat de vierdejaars hogere cijfers behaalden dan de vijfdejaars. Hetzelfde verschijnsel werd aangetroffen bij het analyseren van de gevolgde oplosroutes.
De aanwezigheid van criteriumvaliditeit werd onderzocht door de cijfers die studenten behaalden voor PMP's en PPP's te correleren met hun prestaties op tien cognitieve toetsen. Het bleek niet mogelijk om prestaties op de management problemen te voorspellen uit prestaties op de voor dat doel geselecteerde cognitieve toetsen. De prestaties op cognitieve toetsen onderling bleken wél goed te correleren. Deze constatering leidde tot de conclusie dat de management problemen andere vaardigheden meten dan cognitieve toetsen.
De in hoofdstuk II besproken methode voor het bepalen van de moeilijkheidsgraad van management problemen, werd op geldigheid gecontroleerd door na te gaan of het verschil in moeilijkheidsgraad tussen de management problemen terug gevonden kon worden in

de prestaties. Het als eenvoudig aangeduide PPP werd inderdaad beter opgelost dan het als moeilijk aangeduide PPP. Maar bij de PMP's werden deze verschillen niet aangetroffen. Verondersteld wordt dat de "sturende" aspecten in PMP's (terugkoppeling en cueing) een nivellerende invloed hebben uitgeoefend op het verschil in moeilijkheidsgraad.
Genoemde sturende aspecten worden eveneens verantwoordelijk gehouden voor de hogere prestaties op PMP's in vergelijking met inhoudelijk identieke (en dus even moeilijke) PPP's. Deze constatering leidde tot de conclusie dat PMP's gebruikt kunnen worden als middel om kennis te verwerven.
In tegenstelling tot wat verwacht werd, waren de studenten niet erg enthousiast over het oplossen van PMP's.

Voortbordurend op het thema PMP staat in hoofdstuk IV het zogenaamde CPMP (computerized patient management problem) centraal. Aan de hand van een vergelijking tussen PMP's en CPMP's worden enkele belangrijke kenmerken van laatstgenoemde simulatievorm besproken. Samengevat heeft het gebruik van CPMP's de volgende voordelen boven het gebruik van PMP's:
- responsen kunnen meer tekst bevatten en zijn daardoor vaak realistischer;
- het verschijnsel "cueing" doet zich in mindere mate voor;
- er is geen tijdverlies als gevolg van het noodzakelijke doorbladeren van het testboekje;
- overzichten van reeds verzamelde informatie zijn eenvoudig op te vragen;
- dynamische simulatiemodellen zijn eenvoudiger realiseerbaar;
- oplosroutes zijn achteraf exact reproduceerbaar;
- onmiddellijke, betrouwbare scoring direct na beëindiging van het probleem is mogelijk.

Als nadelen van CPMP's worden genoemd:
- ze zijn minder geschikt als toetsmiddel als niet over een groot aantal microcomputers beschikt kan worden;
- het vervaardigen van de veelal niet uitwisselbare programmatuur is kostbaar;
- het opstellen van scoringsregels is ingewikkeld.

Vervolgens wordt aandacht geschonken aan de constructie van een tandheelkundig CPMP via een bespreking van de vervaardigde software.
Daarna wordt verslag uitgebracht over een pilotstudy naar het functioneren van het CPMP. Het doel van de pilotstudy was het verkrijgen van informatie over het functioneren van de programmatuur en het vernemen van de ervaringen van de deelnemers. Gelet op deze doelstelling werden de volgende conclusies getrokken:
- alle functies in het programmapakket werkten naar bevrediging;
- de deelnemers vonden het programma zeer gebruikersvriendelijk en realistisch;
- toelichting vooraf over de beschikbare functies is noodzakelijk;
- additionele, visuele informatie is gewenst.

Ter illustratie van de analysemogelijkheden van de prestaties op een CPMP worden de oplosprocessen van de deelnemers aan de hand van enkele aspecten (benodigde tijd, hoeveelheid opgevraagde informatie, behandelingsvolgorde) beschreven.

In hoofdstuk V, tenslotte, worden aanbevelingen gedaan voor nieuw onderzoek dat kan leiden tot toepassing op ruimere schaal van PMP's en CPMP's en tot kwaliteitsverbetering van die instrumenten.

SUMMARY

PART I: ASSESSMENT OF PRECLINICAL PERFORMANCE

The central theme of chapter I concerns the important role feedback plays in the acquisition of motor skills. In the training of dental students this feedback, provided by an awareness of results, is of great importance, particularly in the first phase of acquiring motor skills.
"Awareness of results" implies that motor skill performance can be assessed. Generally "work sample tests" are used for this purpose. A work sample test is a replica of a work situation (or part of it), which permits evaluation of the extent to which a particular skill has been acquired. The ultimate purpose in teaching dental students is to ensure they are clinically competent. This competence is an intricate mixture of cognitive, affective and purely motor skills, both qualitatively and quantitatively. The qualitative aspect of motor skills lies in the variety and degree of excellence of the services undertaken. The quantitative aspect is related to the amount of practice required to ensure a constant quality.
If it is possible to give the student a reliable and valid feedback on his performance, the amount of practice may be limited. Thus more time remains to pay attention to other tasks whereby the student can develop in a wider sense and improve his skills.

Chapter II starts with a discussion of the problem inherent in the use of work sample tests, namely subjectivity in judgement. De Groot (1971) mentions five specific problems, which may occur in judging:

1. semantic effect: judgement is influenced by the assessor's opinion of the set task;
2. halo effect: judgement is influenced by significantly good or bad aspects of the task to be performed;
3. sequence effect: judgement is influenced by the quality of previously judged tasks;
4. individual norms: judgement is influenced not only by general but also by personal feelings;
5. contamination effect: the judgemental procedure is, consciously or unconsciously, used for purposes other than the assessing of the task.

It has been shown through studies done in the Netherlands and other countries that such problems affecting objectivity occur in the evaluation of preclinical performance in dental education. In all these studies investigators attempted to improve reliability and validity of preclinical evaluation by concentrating on the following aspects:

- formulation of clear, unequivocal performance criteria;
- establishing a valid system of scoring;
- training of examiners.

In addition studies are reported, carried out in the Department of Dentistry of the Catholic University of Nijmegen, on the evaluation of motor skills. From these the conclusion is drawn that

assessment of work sample tests in preclinical teaching leaves room for much improvement. Arguments are put forward for constructing a new evaluation system.

Chapter III deals with the development of a so called "evaluation protocol". For each important aspect of the Class II cavity preparation the evaluation protocol describes:
- standards for acceptable performance;
- methods for assessing whether the performance meets all the standards;
- means of expressing these observations as a useful score.

The last paragraph deals with the manner whereby the functioning of the evaluation protocol is tested in a pilot study. Both qualitative data (i.e. criticism passed on by both teachers and student assessors) and quantitative data (by which is meant the level of inter and intra-rater agreement) obtained from this study indicated a need for changes in various aspects of the evaluation protocol.

The central theme of chapter IV comprises a description of the development of an individually directed programme of training in evaluation. An attempt is made to improve reliability of evaluation by providing assessors with feedback on the quality of their assessments. Two conditions are discussed, essential to the functioning of feedback:
1. A situation must be created, whereby the potential recipient of feedback will be open to the information received.
2. The information must be offered in such a way as to be of practical use to the recipient.

It is argued that individualised training is more likely to meet the conditions mentioned.
Attention is drawn to two important requirements for individualised training:
- the availability of a sufficient number of student performances;
- a degree of automation of the training programme.

The final paragraph describes the setting up of a study aimed at investigating whether the quality of assessment of preclinical performances might be improved by letting assessors use the evaluation protocol and by training them in evaluating preclinical performance.

Chapter V deals with the psychometric meaning of the concept of "reliability" and the various estimating procedures for this (e.g. "split half", "parallel form"). These procedures are on the whole not suited to determining reliability of judgements. The customary way of estimating this kind of reliability is through an assessment of the inter and intra-rater association or agreement.
The main part of this chapter is taken up with a discussion of the statistical measures used:
- coefficient Kappa: a measure of agreement for nominal data;
- intra class correlation coefficient: a measure of association for interval data;
- index T: a measure of agreement for interval data.

The last paragraph explains what is understood by the validity of a judgement and the method by which this is determined in the present study.

In Chapter VI three questions are introduced. In answering these, results of the study are discussed regarding the functioning of the evaluation protocol and the training programme.
The first question concerns the possible increase in the reliability of preclinical evaluation, if the evaluation protocol is used instead of the present evaluation system. The Kappa coefficients calculated from the judgements made with the evaluation protocol were on average 63 percent larger than the Kappa coefficients calculated from judgements according to the present method. But just as with judgements using the present method, so judgements using the evaluation protocol produced large differences between the student products as measured by inter-rater agreement. The conclusion is drawn that differentiation is necessary between the easily assessed student's results as "clearly satisfactory" or "clearly unsatisfactory" and the much more difficult assessments of "borderline" results. A further conclusion is that even with the evaluation protocol the aspect "finishing of the preparation" cannot be reliably judged.
The second question examines whether preclinical evaluations using the evaluation protocol are more valid than if the present evaluation system had been used. This answer is also in the positive; agreement with the expert score (the standard) was twice as strong with this method (evaluation protocol) than with the present method.
The third question examines whether assessor's judgements improved after partaking in special training sessions. No positive correlation could be found between the amount of training and inter-rater agreement. A positive relationship was found between the amount of training and the time needed for judging by the evaluation protocol.

Finally in chapter VII recommendations are made for the implementation of the evaluation protocol in preclinical teaching and for an experimental study of the effect of training on quality of evaluation.

PART II: ASSESSMENT OF PROBLEM SOLVING ABILITY

Chapter I defines the area on which is focused in the study described in part II of this thesis. After defining "problem" and "problem based learning", the discussion explains that problem solving ability cannot be determined by the fact that a certain problem has been solved. Marshall (1983) offers a better starting point for assessing problem solving ability. According to Marshall competence in problem solving can be measured by the degree of success achieved in providing a satisfactory solution (using a standard considered adequate for the discipline) within an economy of time, at minimal expense and causing the least inconvenience.

Thus problem solving ability can only be achieved if problems are being solved in an efficient way. Ill defined problems can be solved efficiently by using appropriate heuristic procedures. Four general categories of heuristic procedures are discussed:
- trial and error;
- proximity methods;
- fractioning methods;
- knowledge based methods.

A major part of this chapter is taken up with the discussion of the "written simulation" as a method for assessing problem solving ability. Written simulations are descriptions of situations that may confront the practitioner of a particular specialty in practice. The written simulation is assumed to be seen as a real situation by the practitioner. The consequences of his decisions are (on the whole) made known immediately to the person doing the problem solving.
After discussing various forms of written simulation one form in particular is mentioned: the patient management problem (PMP). The basic structure for a PMP is explained, after which is discussed ways in which PMP's can differ from one another. Some drawbacks in the use of PMP's are mentioned. The chapter concludes with reasons for the choice of the PMP as an effective means of instruction and evaluation for a preclinical course in dental education.

Chapter II discusses two methods of effectively learning to establish dental treatment plans. The first method is based on the so called "empirical cycle" and is the one used at present in the dental training. The second method has been recently developed by Verdonschot (1982) and can be distinguished from the first method by the fact it supplies the students with "thinking rules" as well as by the explicit use made of the heuristic procedure "fractioning".
Next defects of the "patient problem on paper" (PPP) are discussed as a method for assessing problem solving ability of dental students. PMP's are argued to be more suitable.
The construction of two dental PMP's is set out. In doing so much attention is paid to the integration of Verdonschot's problem solving model into the structure of the management problems.
Finally the choice of a method for determining the level of difficulty of PMP's is discussed.

In chapter II it is presumed that PMP's might be better suited to the assessing of problem solving ability of dental students than PPP's. In chapter III this presumption leads to a comparison of the validity of the PMP's against the validity of PPP's (contents identical). The types of validity evaluated were specified. In the study two PMP's and two PPP's were examined. PMP1 and PPP1 are based on second year standards and contain an equivalent management problem. PMP2 and PPP2 are also identical in contents but proceed from third year standards. In a compulsory test during the academic year 1983/1984 all third, fourth and fifth year students solved two successive management problems.
By comparing the marks the students of each year got, the existence

of construct validity was investigated. An indication for the existence of construct validity is the finding that fourth and fifth year students achieved higher marks for PMP's and PPP's than third year students did. However the fact that fourth year students obtained higher marks than fifth year students may have diminished the importance of this indication. A similar situation presented itself in analysis of the problem solving pathways.
The existence of criterion validity was investigated by correlating the marks students obtained in PMP's and PPP's with their achievements in ten cognitive tests. It did not appear possible to predict achievement in management problems from achievements in cognitive tests, selected for that purpose. The achievements in the various cognitive tests correlated well however. This led to the conclusion that management problems and cognitive tests measure different things.
The method discussed in chapter II for defining the level of difficulty of management problems was tested by checking if differences in levels of difficulty between the various management problems could be found to occur in the level of achievements. A PPP specified as simple was solved far better than a PPP specified as difficult. These differences were not found with the PMP's. The supposition is that two characteristic aspects of PMP's, namely "feedback" and "cueing", exercised a levelling influence on the difference of levels of difficulty. These aspects were equally held responsible for the higher achievements of PMP's in comparison with PPP's of identical content (i.e. of equal difficulty). This fact led to the conclusion that PMP's can be used as a means to acquire knowledge.
Contrary to expectations students were not too enthusiastic about the solving of PMP's.

Carrying the theme of PMP further leads to CPMP (computerized patient management problem), the main subject of chapter IV. Through a comparison of PMP's and CPMP's some important characteristics of the CPMP type of simulation are discussed. In summary the use of CPMP's has the following advantages over the use of PMP's:
- answers can contain more text, which often makes for greater realism;
- cueing presents itself to a lesser degree;
- time is not lost through having to check back through the test booklet;
- overviews of information already collected are simple to recall;
- dynamic simulation models are simpler to realise;
- problem solving pathways can be reproduced retrospectively;
- immediate and reliable scoring, after finishing the problem, is possible.

Drawbacks to CPMP's are found to be:
- they are less suitable for assessing problem solving ability if large numbers of microcomputers are not available;
- production of software, much of it not interchangeable, is expensive;
- the setting up of a scoring system is complicated.

The construction of a dental CPMP is discussed, taking into account

a survey of the available software. Next a pilot study of the functioning of the CPMP is discussed. The purpose of this pilot study was the gathering of information about the functioning of the software and insight into the views of the participants. With this aim in mind, the following conclusions were drawn:

- all functions of the programme package worked satisfactorily;
- participants found the programme "userfriendly" and realistic;
- information about the available functions in advance is necessary.

Various aspects of the problem solving processes of the participants are described (time needed, amount of information requested, working sequence). These serve to illustrate the possibilities of the CPMP in analysing the students problem solving achievements.

Chapter V contains recommendations for a further study leading to a wider use of PMP's and CPMP's with an improvement in the quality of these instruments.

LITERATUUR

Abou-Rass, M. (1973): A clinical evaluation instrument in endodontics. J. Dent. Educ., 37, 22-36.

Abrams, R.G. en Kelley, M.L. (1974): Student self-evaluation in a pediatric-operative technique course. J. Dent. Educ., 38, 385-391.

Bass, G.M., Moller, J.H. en Johnson, P.E. (1981): New techniques in the construction of patient management problems. Med. Educ., 15, 150-153.

Bennett, E.M., Alpert, R. en Goldstein, A.C. (1954): Communications through limited-response questioning. Public Opinion Quarterly, fall, 303-308.

Berner, E.S., Bligh, T.J. en Guerin, R.O.(1977): An indication for a process dimension in medical problem solving. Med. Educ., 11, 324-328.

Boekaerts, M. (1983): Probleemoplossen: Een eclectische benadering. Tijdschrift voor Onderwijsresearch, 5, 193-217.

Borgesius, T.G. (1973): Beoordelen van practicum werkstukken; prekliniek Conserverende Tandheelkunde. I.O.W.O. 12-73, Katholieke Universiteit, Nijmegen.

Borgesius, T.G. (1980): Beoordelen van practicumwerkstukken: evaluatie van tandheelkundig onderwijsblok 164. I.O.W.O., Katholieke Universiteit, Nijmegen.

Boshuizen, H.P.A. en Claessen, H.F.A.(1982): Cognitieve verwerking en onthouden van patiëntgegevens; een onderzoek bij studenten in Utrecht en Maastricht. In:Schmidt, H.G.(red): Probleemgestuurd onderwijs. Bijdragen tot de onderwijsresearchdagen 1981. Stichting voor Onderzoek van het Onderwijs. Flevodruk Harlingen B.V.

Brennan, R.L. en Prediger, D.J. (1981): Coefficient Kappa: some uses, misuses, and alternatives. Educ. and Psych. Measurement, 41, 687-699.

Briel-van Ingen, T.van den en Plasschaert, A.J.M.(1977): Probleemoplossen in het tandheelkundig onderwijs. Ned. Tijdschr. Tandheelkd., 84, mei, 180-183.

Bruner, J.S. (1970): The growth and structure of skill. In: Connolly, K.L.(Ed.): Mechanisms of skill development. Academic Press, New York. P63-94.

Buis, P. (1978): Het functioneren van terugkoppeling in het wetenschappelijk onderwijs: Twee voorafgaande voorwaarden. Swets & Zeitlinger B.V., Amsterdam/Lisse.

Campbell, D.T. en Stanley, J.C. (1971): Experimental and quasi-experimental designs for research on teaching. In: Gage, N. L.(Ed.), Handbook of research on teaching, Rand McNally & Company, Chicago.

Cassidy, R.E., Marshall, F.J., Gaston, G.W. en Snodgrass, M. (1972): Computer-assisted instruction for diagnostic problem solving of toothache. J. Dent. Educ., 36, 46-56.

Cohen, J. (1960): A coefficient of agreement for nominal scales. Educ. and Psych. Measurement, 20, 37-46.

Cohen, S.N. en Silvestri, A.R. (1980): Evaluation of describing and grading preclinical technical exercises. J.Dent.Educ. 44, 547-549.

Corte, E. de, Geerligs, C.T., Lagerweij, N.A.J., Peters, J.J. en Vandenberghe, R.(1976): Beknopte didaxologie. Wolters-Noordhoff, Groningen.

Crombag, H.F.M. (1973): Het oefenen van vaardigheden: het juridisch practicum. In: Woerden, W.M. van, Chang, T.M. en Geuns-Wiegman, L.J.M. van (red): Onderwijs in de maak. Uitgeverij Het Spectrum, Utrecht/Antwerpen.

Drenth, P.J.D. (1973): De psychologische test: Een inleiding in de theorie van de psychologische test en zijn toepassingen. Van Loghum Slaterus, Deventer.

Edwards, W.S., Morse, P.K. en Mitchell, R.J. (1982): A practical evaluation system for preclinical restorative dentistry. J. Dent. Educ., 46, 693-696.

Elstein, A.S., Shulman, L.S. en Sprafka, S.A. (1978): Medical problem solving: An analysis of clinical reasoning. Harvard University Press, Cambridge, Massachusetts.

Fitts, P.M. (1964): Perceptual-motor skill learning. In: Melton, A.W.(Ed.): Categories of human learning. Academic Press, New York. P253-265.

Fitts, P.M., Bahrick, H.P., Noble, M.E. en Briggs, G.E. (1959): Skilled performance. United States Air Force, Wright Air Development Centre, Final Report, No. AF41: 657-670.

Fitts, P.M. en Posner, M.I. (1967): Human performance. Belmont, Wadsworth.

Fitzpatrick, R. en Morrison, E.J. (1971): Performance and product evaluation. In: Thorndike, R.L. (ed.): Educational measurement. American Council on Education, 1971.

Fleiss, J.L., Cohen, J. en Everitt, B.S. (1969): Large sample standard errors of Kappa and weighted Kappa. Psych. Bull., 72, 323-327.

Frijda, N.H. en Elshout, J.J. (1976): Probleemoplossen en denken. In: Michon, J.A., Eykman, E.G.J. en Klerk, L.F.W. de (red): Handboek der Psychonomie. Van Loghum Slaterus, Deventer.

Fuller, J.L. (1972): The effects of training and criterion models on interjudge reliability. J. Dent. Educ., 36, 19-22.

Gaines, W.G., Rasmussen, R.H. en Uchello, E. (1975): Increasing the objectivity of clinical grading. Dent. Hyg.,49, 227-280.

Geissler, P.R. (1973): Student self-assessment in dental technology. J. Dent. Educ., 37, 19-21.

Goran, M.J., Williamson, J.W. en Gonnella, J.S.(1973): The validity of patient management problems. J. Med. Educ., 55, 529-537.

Graaff, E. de en Galesloot, J.A.M. (1982): De ontwikkeling van een toetsmethode voor "medisch probleemoplossen". In: Schmidt, H. G.(red): Probleemgestuurd onderwijs. Bijdragen tot de Onderwijsresearchdagen 1981. Stichting voor Onderzoek van het Onderwijs, Flevodruk Harlingen B.V.

Graaff, E. de, Moust, J.H.C., Ronteltap, C.F.M. en Schmidt, H.G. (1982): Studiebeleving van Maastrichtse medische studenten. In: Schmidt, H.G.(red): Probleemgestuurd Onderwijs. Bijdragen tot de Onderwijsresearchdagen 1981.Stichting voor Onderzoek van het Onderwijs. Flevodruk Harlingen B.V.

Greeno, J.G. (1980): Trends in the theory of knowledge to problem solving. In: Tuma, D.T. en Reif, F.(ed): Problem solving and education: Issues in teaching and research. Lawrence Erlbaum Associates, Publishers, Hillsdale, New Jersey.

Groot, A.D. de (1971): Methodologie: Grondslagen van onderzoek en denken in de gedragswetenschappen. Mouton & Co, 's-Gravenhage.

Groot, A.D. de en Naerssen, R.F. van (1975): Studietoetsen; construeren, afnemen, analyseren. Deel II. Mouton, Den Haag.

Groot, A.D. de (1983): Is de kwaliteit van het onderwijs te beoordelen? In: De kwaliteit van het onderwijs. Een bundel naar aanleiding van een onderwijscongres gehouden ter gelegenheid van het tienjarig bestaan van het RION (Research Instituut voor het Onderwijs in het Noorden). Wolters-Noordhoff, Groningen.

Guilford, J.P. en Fruchter, B. (1973): Fundamental statistics in Psychology and Education. Tokyo.

Harden, R.M. (1983): Preparation and presentation of patient management problems (PMPs). Med. Educ., 17, 256-276.

Hartmann, D.P. (1977): Considerations in the choice of inter-observer reliability estimates. J. applied behav. analysis, 10, 103-116.

Hasman, A. (1983): Computer geassisteerde medische diagnostiek. VMBI-mededelingen, 12, 2, 29-33.

Hayes, J.R. (1981): The complete problem solver. The Franklin Institute Press, Philadelphia.

Hinkelman, K.W. en Long, N.K. (1973): Method for decreasing subjective evaluation in preclinical restorative dentistry. J. Dent. Educ., 37, 13-18.

Hoffer, E.P., Barnett, G.O., Farquar, B.B. en Prather, P.A. (1975): Computer-aided instruction in medicine. Annual Review of Biophysics and Bio engineering, 4, 103.

Houpt, M.I. en Kress, G. (1973): Accuracy of measurement of clinical performance in dentistry. J. Dent. Educ., 37, 34-46.

Hyman, J.J. en Doblecki, W. (1983): Computerized endodontic diagnosis. JADA, Vol.107, 755-758.

Instituut Conserverende Tandheelkunde voor Volwassenen (1978): Syllabus klinische Tandheelkunde. Katholieke Universiteit, Nijmegen.

Instituut Conserverende Tandheelkunde voor Volwassenen (1982): Syllabus bij de geïndividualiseerde kursus Preparatie /Restauratie I en II, Blok 155, Blok 255. Katholieke Universiteit, Nijmegen.

Instituut Conserverende Tandheelkunde voor Volwassenen (1983): Handleiding blok 155: Preparatie/Restauratie. Katholieke Universiteit, Nijmegen

Johnson, D.M. (1972): A systematic introduction to the psychology of thinking. Harper & Row, New York.

Jong, T. de en Ferguson-Hessler, M.G.M. (1982): Voorwaarden voor het succesvol oplossen van problemen. Groep Onderwijsresearch, rapport no. 30, T.H. Eindhoven.

Jong, T. de en Ferguson-Hessler, M.G.M. (1984): Strategiegebruik bij het oplossen van problemen in een semantisch rijk domein: electriciteit en magnetisme. Tijdschrift voor Onderwijsresearch, 9,1,3-15.

Käyser, A.F. (1981): Probleemoplossing en behandelingsplan. In: Handboek voor de tandheelkundige praktijk, A3.I-3. Bohn, Scheltema & Holkema, Utrecht/Antwerpen.

Keele, S.W. (1968): Movement control in skilled motor performance. Psych. Bull., 70, 387-403.

Kendall, M.G. en Stuart, A. (1961): The advanced theory of statistics Vol.II. Griffin, London.

Landis, J.R. en Koch, G.G. (1975): A review of statistical methods in the analysis of data arising from observer reliability studies (Part II). Statistica Neerlandica, nr 4, 151-161.

Lawlis, G.F. en Lu, E. (1972): Judgment of counseling process: Reliability, agreement, and error. Psych. Bull., 78, 17-20.

Lenat, D.B. (1984): Computer software for intelligent systems. Scientific American, Vol. 251, 3, 152-160.

Light, R.J. (1973): Issues in the analysis of qualitative data. In: Travers, R.M.W. (Ed.): Second handbook of research on teaching. Rand McNally College Publishing Company, Chicago.

Lu, K.H. (1971): A measure of agreement among subjective judgments. Educ. and Psych. Measurement, 31, 75-84.

Mackenzie, R.S. (1973): Defining clinical competence in terms of quality, quantity, and need for performance criteria. J. Dent. Educ. 37: 37-44.

Mackenzie, R.S. (1974): Factors essential to evaluation of clinical performance. J. Dent. Educ., 38, 214-223.

Marquis, Y., Chaoulli, J., Bordage, G., Chabot, J.M. en Leclere, H. (1984): Patient management problems as a learning tool for the continuing medical education of general practitioners. Med. Educ., 18, 117-124.

Marshall, J.R. (1977): Assessment of problem solving ability. Med. Educ., 11, 171-177.

Marshall, J.R. (1983): How we measure problem solving ability. Med. Educ., 17, 319-324.

Massler, M. en Evans, J. (1977): Correlation between preclinical and clinical grades. J. Dent. Educ. 41, 596-570

McGuire, C.H., Solomon, L.M. en Bashook, P.G. (1976): Construction and use of written simulations. The Psychological Corporation, New York.

Mettes, C.T.C.W. en Pilot, A. (1980): Onderwijs in het oplossen van vraagstukken. Onderwijskundig Centrum CDO/AVC, Bulletin 14, Technische Hogeschool Twente.

Metz, J.C.M. (1984): Medische competentie: Een onderzoek naar de betrouwbaarheid en validiteit van het gestructureerd klinisch examen. Proefschrift, Katholieke Universiteit, Nijmegen.

Millman, J. (1974): Criterion referenced measurement. In: Popham, W.J.(ed): Evaluation in education. Englewood Cliffs, New Jersey.

Mitchell, S.K. (1979): Interobserver agreement, reliability, and generalizability of data collected in observational studies. Psych. Bull., 86, 376-390.

Moonen, J. en Gastkemper, F. (1983): Computergestuurd onderwijs. In de reeks: "Onderwijskundige informatie voor het Hoger Onderwijs". Aula pocket 811. Uitgeverij Het Spectrum, Utrecht/Antwerpen.

Mulder, G., Michon, J.A. en Moraal, J. (1976): Motorische vaardigheden. In: Michon, J.A., Eijkman, E.G.J. en Klerk, L.F.W. de (Red.): Handboek der Psychonomie. Van Loghum Slaterus, Deventer.

Mullaney, T.P., Smith, T.A., Duell, R.C. en Kaplan, A. (1976): Four-phase study of computer-assisted and slide-tape methods of simulating clinical endodontic problems. J. Dent. Educ., 40, 681-687.

Natkin, E. en Guild, R.E. (1967): Evaluation of preclinical laboratory performance: a systematic study. J. Dent. Educ., 31, 152-161.

Newble, D.I., Hoare, J. en Baxter, A. (1982): Patient Management Problems: issues of validity. Med. Educ., 16, 137-142.

Newell, A. en Simon, H.A. (1972): Human problem solving. Englewood Cliffs, New Jersey, Prentice Hall Inc.

Norman, G.R. en Feightner, J.W. (1981): A comparison of behaviour on simulated patients and patient management problems. Med. Educ., 15, 26-32.

Oers, H.J.M.van (1981): Zelfstandige kennisverwerving als aspect van het wetenschappelijk denken. Ped. Studiën, 58, 420-432.

Otto, F.L. (1979a): Evaluatie beoordelingsprocedure blok 160c 1976-1977. Een deelrapport van de Subcommissie Toetsing en Beoordeling Motorische Vaardigheden. Katholieke Universiteit, Nijmegen.

Otto, F.L. (1979b): Evaluatie beoordelingsprocedure blok 152 1977-1978. Een rapport ten behoeve van de Subcommissie Toetsing en Beoordeling Motorische Vaardigheden. Katholieke Universiteit, Nijmegen.

Otto, F.L. (1981): Beoordelingsprocedures beoordeeld: Een systematische evaluatie van beoordelingsprocedures van de motorische onderwijsblokken eerste cursusjaar Tandheelkunde. Katholieke Universiteit, Nijmegen.

Page, G.G. en Fielding, D.W. (1980): Performance on PMPs and performance in practice: Are they related? J. Med. Educ., 55, 529-537.

Patridge, M.I. en Mast, T.A. (1978): Dental clinical evaluation: A review of the research. J. Dent. Educ., 42, 300-305.

Povenmire, H.K. en Roscoe, S.N. (1971): An evaluation of ground-based flight trainers in routine primary flight training. Human Factors, 13, 109-116.

Pryor, H.G. en Racey, G. (1982): Minicomputer simulation of medical emergencies and advanced life support. J. Dent. Educ., 46, 657-660.

Reynolds, H.T. (1977): Analysis of nominal data. Sage University Paper series on Quantitative Applications in the Social Sciences, 07-007. Sage Publications, Beverly Hills and London.

Rimoldi, H.J.A. (1961): The test of diagnostic skills. J. Med. Educ., 36, 73-79.

Roscoe, S.N. (1971): Incremental transfer effectiveness. Human Factors, 13, 561-567.

Roscoe, S.N. (1972): A little more on incremental transfer effectiveness. Human Factors, 14, 363-364.

Ryge, G. en Snyder, M. (1973): Evaluating the clinical quality of restorations. J. Am. Dent. Assoc., 87, 369-377.

Salvendy, G., Hinton, W.M., Ferguson, G.W. en Cunningham, P.R. (1973): Pilot study on criteria in cavity preparation- facts or artifacts? J. Dent. Educ. 37: 27-31.

Salvendy, G., Joost, M.G., Cunningham, P.R., Ferguson, G.W., Wilko, R.A. en Dees, R.W. (1976): Improving evaluation of amalgam restorations. J. Dent. Educ., 40, 6, 368-369.

Sanders, A.J. (1980): Evaluatierapport blok 155 studiejaar 1977-1978. Instituut Conserverende Tandheelkunde voor Volwassenen, Katholieke Universiteit, Nijmegen.

Sanders, A.J. (1984): STAPAM: een programmapakket voor het simuleren van patiënt management in de tandheelkunde. Intern Rapport CE 84-.., Instituut Conserverende Tandheelkunde voor Volwassenen, Katholieke Universiteit, Nijmegen.

Sanders, A.J. en Kortsmit, W.J.P.M. (1983): Een programma voor het berekenen van overeenstemming tussen beoordelingsscores. Intern Rapport CE-.... Instituut Conserverende Tandheelkunde voor Volwassenen, Katholieke Universiteit, Nijmegen.

Sanders, A.J. en Straetmans, G.J.J.M. (1982): Beschrijvingen van beoordelingsprotocollen klasse II-preparatie. Instituut Conserverende Tandheelkunde voor Volwassenen, Katholieke Universiteit, Nijmegen.

Schiff, A.J., Salvendy, G., Root, C.M., Ferguson, G.W. en Cunningham, P.R. (1975): Objective evaluation of quality in cavity preparations. J. Dent. Educ., 39, 2, 92-96.

Schmidt, H.G. (1979): Leren met problemen. Een inleiding in probleemgestuurd onderwijs. In: Handboek voor de onderwijspraktijk, deel III, sectie 3.4.

Schmidt, H.G. (1982): Enkele cognitieve effecten van probleemgestuurd onderwijs. In: Schmidt, H.G.(red): Probleemgestuurd Onderwijs. Bijdragen tot de Onderwijsresearchdagen 1981. Stichting voor Onderzoek van het Onderwijs, Flevodruk Harlingen B.V.

Schmidt, R.A. (1975): A schema theory of discrete motor skill learning. Psych. Review, Vol. 82, 4.

Schwartz, M.W. en Hanson, C.W. (1982): Microcomputers and computer-based instruction. J. Med. Educ., 57, 303-307.

Scott, W.A. (1955): Reliability of content analysis: The case of nominal scale coding. Public Opinion Quarterly, 19,321-325.

Shortliffe, E.H. (1976): Computer based medical consultations MYCIN. American Elsevier, New York.

Shrout, P.E. en Fleiss, J.L. (1979): Intraclass correlations: Uses in assessing rater reliability. Psych. Bull. 86, 2, 420-428.

Silvestri, A.R., Cohen, S.N. en Singh, I. (1979): The improvement of technical skills in preclinical courses. J.Dent.Educ. 43, 641-644.

Sokolow, S. en Solberg, W. (1971): Computer-assisted instruction in dental diagnosis: a product development. J. Dent. Educ., 35, 349-355.

Steures, R.W.R. en Tromp, Th.J.M. (1978): Vernieuwing van een practicum voor tandheelkundige handvaardigheden, deel I, II en III. Ned. Tijdschr. Tandheelkd., 85, 421-426; 87, 225- 230, 258-264.

Straetmans, G.J.J.M. (1982): Het onderwijsstimuleringsproject 'Een geïndividualiseerd trainingsprogramma voor het beoordelen van praktikumwerkstukken'. Intern rapport CE 82-04. Katholieke Universiteit, Nijmegen.

Straetmans, G.J.J.M. en Sanders, A.J. (1984): Enkele associatie- en overeenstemmingsmaten voor beoordelingsscores op nominaal en interval meetniveau. Intern Rapport CE 84-.., Instituut Conserverende Tandheelkunde voor Volwassenen, Katholieke Universiteit, Nijmegen.

Studiegids Tandheelkunde 1984-1985. Katholieke Universiteit, Nijmegen.

Taylor, W.C., Grace, M., Taylor, T.R., Fincham, S.M. en Skakun, E.N. (1976): The use of computerized patient management problems in a certifying examination. Med. Educ., 10, 179-182.

Tinsley, H.E.A. en Weiss, D.J. (1975): Interrater reliability and agreement of subjective judgments. J. of Counseling Psych., 22, 358-376.

Vaags, D.W. (1975): Over het oplossen van technische problemen. Proefschrift, Technische Hogeschool, Eindhoven.

Verbeek, H.A. (1982): Modernisering van klinisch onderwijs. Ned. Tijdschr. Geneeskd., 126, 39, 1787-1790.

Verdonschot, E.H.A.M. (1980): Beoordeling van tandheelkundige behandelplannen. Ned. Tijdschr. Tandheelkd., 87, 432-438.

Verdonschot, E.H.A.M. (1982): Een probleemoplossingsmodel voor complexe tandheelkundige vraagstukken. Ned. Tijdschr. Tandheelkd., 89, 405-414.

Verdonschot, E.H.A.M. (1983): Geprogrammeerde instructie ten behoeve van het opstellen van een PMP. Intern Rapport CE 83-01. Instituut Conserverende Tandheelkunde voor Volwassenen, Katholieke Universiteit, Nijmegen.

Verdonschot, E.H.A.M. (1984): Dental treatment planning and problem solving. Proefschrift, Katholieke Universiteit, Nijmegen.

Willems, J. (1978): Probleemgeoriënteerd (groeps)onderwijs: een methode om kennis te laten functioneren. I.O.W.O., Katholieke Universiteit, Nijmegen.

Wilson, C.L. (1965): On-the-job and operational criteria. In: Glaser, R. (ed): Training research and education. John Wiley & Sons, Inc., New York.

Zeeuw, J. de (1978): Algemene Psychodiagnostiek II: testtheorie. Swets & Zeitlinger B.V., Amsterdam.

BIJLAGEN

BIJLAGE 1

BEOORDELINGSPROTOCOL TEN BEHOEVE VAN DE KLASSE II-TWEEVLAKSPREPARATIE VOOR AMALGAAM. OVERZICHT VAN DE BEOORDELINGSASPECTEN.

1. outline	1.1 hoofdfissuurpatroon
	1.2 preparatiebreedte
	1.3 approximale wand
	1.4 buccale wand
	1.5 palatinale/linguale wand
	1.6 buccale wand box
	1.7 palatinale/linguale wand box
	1.8 cervicale wand box
2. diepte	2.1 step mesiaal
	2.2 step distaal
	2.3 richting bodem
	2.4 diepte box axiaal
	2.5 richting axiale wand
3. cav. oppervlakte hoek	3.1 step buccaal
	3.2 step palatinaal/linguaal
	3.3 box buccaal
	3.4 box palatinaal/linguaal
	3.5 box cervicaal
4. convergentie	4.1 approximale wand
	4.2 buccale wand step
	4.3 palatinale/linguale wand step
	4.4 convergentie zijfissuur
	4.5 buccale wand box
	4.6 palatinale/linguale wand box
5. pulpo ax. afschuining	5.1 breedte vlak
6. afwerking preparatie	6.1 outline step
	6.2 outline box
	6.3 bodem step
	6.4 bodem box
	6.5 buccale wand
	6.6 palatinale/linguale wand
	6.7 axiale wand

BIJLAGE 2

COMPUTERVERWERKING VAN DE TRAININGSGEGEVENS

Een belangrijke doelstelling van de geautomatiseerde administratie van de trainings-gegevens is het snel kunnen terugkoppelen naar de beoordelaars over hun prestaties. Hierbij is het in de eerste plaats nodig elke individuele beoordelaar snel te informeren over zijn prestatie. Daarnaast moeten snel gegevens beschikbaar komen over de prestaties van alle aan de training deelnemende personen in de gegeven situatie. Deze gegevens zouden positief kunnen bijdragen aan de kwaliteit van toekomstige beoordelingen. Een randvoorwaarde waaraan een trainings-systeem in de gegeven omstandigheden zeker moet voldoen, is de mogelijkheid om de deelnemers onafhankelijk van elkaar op een voor ieder geschikte plaats en tijdstip te kunnen laten werken. Hierin kan worden voorzien door te werken met een verrijdbaar computersysteem dat ter plekke de training ondersteunt. Als tevens voorzien kan worden in een eenvoudige bediening van de apparatuur, kan in feite de trainings-sessie zelfstandig worden uitgevoerd. Het gebruik van de computer heeft als bijkomend voordeel dat alle voor een beoordelaar op een bepaald moment niet relevante beoordelingsinformatie afgeschermd kan worden.

De opbouw van het programmapakket.

Ter beperking van de hoeveelheid zelf te ontwikkelen programmatuur werd aansluiting gezocht bij een database-programmapakket zoals dat op de aanwezige apparatuur (microcomputer) beschikbaar was. Het database-pakket is eenvoudig van opzet, maar toereikend voor de eisen van het trainingsprogramma. Met het pakket zijn standaard de volgende functies mogelijk:

- het definiëren van een gegevensbestand (werkstuknummer, beoordelaarsidentificatie, datum, kenmerken, etc.);
- het wijzigen van gegevens in het bestand;
- het op een beeldscherm of papier afdrukken van (een deel van) de inhoud van het bestand;
- het selecteren en creëren van een deelbestand;
- het sorteren van het bestand;
- het met de hand invoeren van nieuwe gegevens in het bestand;
- het informeren van de gebruiker (een "help"-functie).

Aan deze functies werden toegevoegd:

a. een programma om beoordelingen vanaf een schrapkaart te lezen, de gegevens toe te voegen aan het bestand en de beoordelaar te informeren over de overeenstemming van het eigen oordeel met een referentie-oordeel en met de oordelen van collega's.
b. een programma om meer globale kwaliteitsgegevens te verkrijgen in de vorm van Kappa-coëfficiënten.

ad a. Terugkoppeling van beoordelingsgegevens.

Om de snelheid van terugkoppeling maximaal te laten zijn is het programma voor de verwerking van schrapkaarten in machinetaal (Z80-code) geschreven. De gebruiker kan via een eenvoudige dialoog met het programma opgeven welke in- en uitvoer gewenst wordt, en met welke invoerdatum de beoordelingen in het bestand moeten worden bijgeschreven. Het verwerkingsprogramma maakt gebruik van een hulpbestand dat de volgende onderdelen bevat:
- de definitie van het beoordelingenbestand. Met deze definitie kan getest worden of het in te lezen beoordelingenbestand de juiste opbouw heeft.
- definities van schrapkaartbeelden. In principe kunnen meerdere soorten kaarten aangeboden worden. Elk kaarttype wordt hierbij op de kaarten zelf aangeduid met een (vaste) kaartcode.
- de koppeling van schrapkaartkolommen met de recordvelden uit het beoordelingenbestand.
- de relatie van een recordveld (een beoordelingskenmerk) tot andere velden en de naam van het veld. De naam van het veld wordt voor de uitvoer gebruikt. De relatie wordt aangegeven door te vermelden of er subkenmerkvelden bestaan. Deze informatie wordt gebruikt om bij de invoer van kenmerkbeoordelingen een overzicht te geven van nader te beoordelen subkenmerken.
- de naam van het beoordelingenbestand.

Het hulpbestand wordt samengesteld aan de hand van de definitie van een beoordelingenbestand. Voor het creëren bestaat geen apart programma, het hulpbestand moet handmatig worden opgebouwd.

De snelheid van het programma wordt nog bevorderd door de eis dat het beoordelingenbestand volledig in het computergeheugen moet kunnen worden opgenomen. Tijdens het gebruik echter groeit het bestand. Als het bestand te groot wordt moet een nieuw bestand worden gemaakt met behulp van de standaard selectie-functie van het database-pakket. Deze werkwijze voorziet tevens in het up-to-date houden van het beoordelingenbestand. In het algemeen immers zullen de oordelen "verouderen" en zijn slechts de meest recente oordelen van collega's interessant. Fouten die tijdens het gebruik in het beoordelingenbestand ontstaan kunnen met de wijzigingsfunctie van het database-pakket ongedaan gemaakt worden.

ad b. Het bepalen van de overeenstemming.

Coëfficiënt Kappa is een maat voor de overeenstemming. Deze coëfficiënt kan op meerdere manieren over beoordelingsparen berekend worden. Het ontwikkelde programma berekent een aantal Kappa-versies. De invoer van het programma is een (deel)verzameling van oordelen uit het beoordelingenbestand, die met de selectie-functie van het database-pakket gemaakt kan worden. Het bestand dat verkregen wordt, bevat alleen afdrukbare karakters en kan met de tekstverwerker van de microcomputer verder bewerkt worden (bijvoorbeeld het koppelen van bestanden). Het programma is geschreven in Microsoft BASIC en wordt in de gecompileerde vorm gebruikt. In een dialoog bepaalt de gebruiker de werkstukken, de

beoordelingsdata, de (sub)kenmerken, en de beoordelaarsparen waarover Kappa's berekend worden. De uitvoer vindt plaats op papier en bestaat uit een aanduiding van de beoordelaarsparen, de werkstukidentificaties, de kenmerken, een frequentietabel van de gegeven beoordelingen, het overeenstemmingspercentage en enkele Kappa-waarden (zie figuur 4.1 in deel I).

BIJLAGE 3

DE PLENAIRE NABESPREKINGEN

In elke plenaire nabespreking werden aan de deelnemers overzichten verstrekt van de door hen geleverde beoordelingsprestaties in de voorafgaande trainings-sessie. Aan de hand daarvan werden "probleemgevallen" besproken. Bijvoorbeeld:
- werkstukken waarover erg veel meningsverschillen waren.
- (sub)kenmerken waarvoor overeenstemming niet of nauwelijks bereikt werd.

Bij deze besprekingen werd geprobeerd om de redenen te achterhalen van de lage overeenstemmingen en op grond van overleg te komen tot een oplossing. De belangrijkste constateringen tijdens deze besprekingen waren:

Met betrekking tot het beoordelingsprotocol.
- Subkenmerk 1.5 is alleen van toepassing op ondermolaren.
- Subkenmerk 5.1 betreft een aspect dat in veel gevallen onzichtbaar is. Gesuggereerd werd om de driepunts-schaal voor dit subkenmerk te vervangen door een tweepunts-schaal: 2 = aanwezig; 1 = afwezig.
- Bij de beoordeling van werkstukken op subkenmerk 4.2 en 4.3 moet rekening gehouden worden met de inzetrichting van de boor.

Met betrekking tot de werkstukken.
- Bij nog al wat werkstukken werd kritiek geleverd op de positionering van het buurelement. Een niet juiste positionering heeft tot gevolg dat de omschrijvingen in het beoordelingsprotocol in sommige gevallen niet bruikbaar zijn. Voorgesteld werd om werkstukken in een fantoomkaak op te stellen, waardoor de positionering van de buurelementen automatisch juist is. Daarnaast levert dit het voordeel op dat de werkelijke beoordelingssituatie beter benaderd wordt. In de prekliniek worden de werkstukken ook in de kaak beoordeeld.
- Bij sommige werkstukken waren de beoordelaars het absoluut oneens met het referentie-oordeel.

Met betrekking tot de trainings-sessies zelf.
- Sommige deelnemers vonden dat de trainings-sessies te lang duurden waardoor de concentratie verslapte.
- Sommige deelnemers zagen meer in de beoordelingsvolgorde "subkenmerk-kenmerk" dan in de omgekeerde volgorde. Door eerst op subkenmerkniveau te beoordelen zou men beter in staat zijn om een accuraat kenmerk-oordeel te geven.

Met betrekking tot de plenaire nabesprekingen.
- De plenaire nabesprekingen werden als zinnig ervaren. Unaniem was men van mening, dat door "probleemgevallen" ter discussie te stellen de aandacht nog eens gevestigd werd op moeilijke beoordelingssituaties, hetgeen voor de komende trainings-sessies (en voor het beoordelen in de prekliniek) nuttig zou kunnen zijn.

BIJLAGE 4

OVERZICHT VAN DE KENMERK- EN SUBKENMERKBEOORDELINGEN VAN 4 BEOORDELAARS OP 6 WERKSTUKKEN (DIRECTE VERGELIJKING).

werks	ass	O	D	C	C'	P	A	O	D	C	C'	P	A
36	1	1	3	3	2	2	1	22299312	32212	22222	222922	2	3222122
36	2	2	2	2	3	1	1	22299322	22212	22222	222922	2	3222322
36	4	2	3	3	2	3	2	12299312	22232	32222	222922	2	2232222
36	5	2	2	2	1	1	2	22299312	22212	32222	222922	1	2222222
128	1	1	3	1	3	2	1	22299212	22222	22212	222922	1	2212222
128	2	2	3	2	3	3	2	22299111	22222	22212	222922	2	3212332
128	4	1	3	2	2	2	2	22399111	22212	22322	321921	2	2212222
128	5	1	2	1	2	1	2	22299111	22222	22212	222922	2	2212222
374	1	1	2	1	1	3	2	22119112	23222	32232	232232	2	2222221
374	2	1	2	2	1	3	2	22119111	22222	32312	332232	2	2322223
374	4	1	2	1	1	2	2	22129111	22222	32232	232232	2	2112212
374	5	1	1	1	1	2	2	22119111	22122	22232	232232	2	2122222
449	1	1	2	3	2	2	2	22119112	32231	22222	222322	2	2212212
449	2	1	3	3	3	3	2	22119112	22222	22222	222222	2	2212122
449	4	1	2	3	1	3	1	22129111	22221	22232	322322	2	2212222
449	5	1	2	2	1	3	1	22119111	22221	22232	322322	2	2212222
658	1	1	2	2	2	3	2	22112212	22221	22222	222323	2	2212222
658	2	1	2	3	2	2	2	12111222	23222	22222	222223	2	2212112
658	4	3	3	3	2	3	2	22111212	22221	22222	323223	2	2212222
658	5	1	3	3	2	2	2	22112112	22221	22222	222322	1	2212222
868	1	1	3	2	2	3	2	22299212	22231	22222	223922	2	3211222
868	2	2	2	2	3	3	2	22299212	22222	22222	222922	2	3313332
868	4	2	3	2	2	2	2	22299212	22222	22322	222922	2	2222322
868	5	2	1	3	3	2	1	22299212	22222	22222	223922	2	3322322

O = outline; D = diepte; C = caviteit oppervlakte hoek; C' = convergentie; P = pulpo-axiale afschuining; A = afwerking; werks = identificatienummer van het werkstuk; ass = assistentnummer.

BIJLAGE 5

BETROUWBAARHEID VAN CIJFERS GEBASEERD OP KENMERKSCORES VERSUS IMPRESSIONISTISCHE (GLANCE-AND-GRADE) CIJFERS

Cijfers gebaseerd op kenmerkscores

De cijfers die in het preklinisch onderwijs aan de klasse II-tweevlakspreparaties worden toegekend zijn transformaties van gesommeerde kenmerkscores. De transformatieregel die gebruikt wordt luidt als volgt:

	cijfer									
	1	2	3	4	5	6	7	8	9	10
gesommeerde kenmerkscores	6	7	8	9/10	11/12	13	14/15	16	17	18

Toepassing van deze regel op de kenmerkscores van de in de trainings-sessies aangeboden werkstukken, leverde de volgende cijfers op:

	beoordelaar						beoordelaar						beoordelaar				
werk stuk	1	2	3	4	5	werk stuk	1	2	3	4	5	werk stuk	1	2	3	4	5
4	4	4	4	4	5	429	4	3	1	4	4	779	4	3	4	3	4
36	5	5	6	7	4	449	5	4	4	5	5	785	5	5	4	5	3
45	2	4	3	2	3		5	7	4	5	4	806	5	8	7	6	7
128	5	7	6	5	4	485	5	5	4	5	4	819	5	3	2	4	4
	5	7	4	6	4	498	4	4	4	5	2	868	6	7	6	6	5
202	2	4	3	3	1	537	5	6	5	5	5	869	5	5	5	7	5
229	4	4	2	3	3	541	5	4	6	5	7		5	5	5	6	7
257	4	7	7	5	5	644	6	7	7	7	5	893	7	6	5	7	7
278	4	4	4	4	4	658	5	5	5	8	6	897	4	6	5	6	6
288	4	4	1	4	1	728	4	5	5	3	3	931	5	5	4	7	6
305	5	5	5	7	5		3	4	4	3	3	954	7	7	7	7	7
	5	7	8	7	6	734	4	4	6	5	4		7	8	5	5	7
312	4	4	5	6	3	746	2	4	4	5	2	957	4	4	3	3	4
364	5	5	8	9	6		3	5	2	4	3	993	7	8	7	5	7
374	4	5	5	4	3	767	7	7	5	6	5	994	5	5	4	4	5
413	3	7	6	6	5		7	8	8	7	6						

De betrouwbaarheid van de berekende cijfers wordt geschat via de intraklasse correlatie coëfficiënt (een associatiemaat) en via index T (een overeenstemmingsmaat). Zie par. 5.3.3 en 5.3.4 in deel I voor een bespreking van deze statistische maten.

De intraklasse correlatie coëfficiënt (ICC)

De ICC wordt geschat op basis van de variantie-schattingen afkomstig uit een variantie-analyse. Voor het onderhavige geval is een tweeweg variantie-analyse (mixed model) vereist.

tweeweg variantie analyse (mixed model)

bron	SS	df	MS	F
werkstukken	357.54	46	7.77	
beoordelaars	21.60	4	5.40	5.63*
error	176.80	184	0.96	
totaal	555.94	234		

* $p<.01$

De ICC wordt gegeven door:

$R = (MS_w - MS_e)/(MS_w + (k-1)MS_e) = 0.59$. Toetsing voor significantie van R vindt plaats aan de hand van de gevonden F-waarde voor het werkstukkeneffect. Bij 46 en 184 vrijheidsgraden voor respectievelijk teller en noemer is de resulterende F-waarde significant op het 1-procent toetsingsniveau. De gevonden R geeft aan dat de betrouwbaarheid van het oordeel van één enkele beoordelaar erg matig is. Het lijkt dus niet verstandig om zak-/slaagbeslissingen te nemen op basis van het oordeel van één beoordelaar.

index T

Index T, een overeenstemmingsmaat voor data van interval niveau, wordt berekend als na uitvoering van een nonparametrische chikwadraat toets is gebleken, dat de overeenstemming niet toegeschreven kan worden aan het toeval alleen. Als criterium voor overeenstemming geldt dat cijfers van beoordelaars niet meer dan één punt van elkaar mogen verschillen. In 10 gevallen zijn de beoordelaars het met elkaar eens. De kans dat de vijf beoordelaars toevallig overeenstemming bereiken is: $p=(n-1)\sum 2+n/n^k = 0.0028$;

waarbij n = aantal schaalpunten en k = aantal beoordelaars. χ^2=669.212. De steekproefverdeling van deze chi-kwadraat is een chi-kwadraatverdeling met één vrijheidsgraad. De nulhypothese dat de overeenstemmingen toevallig zijn wordt verworpen (p<.001). Vervolgens kan de mate van overeenstemming berekend worden met index T. T corrigeert de gevonden overeenstemming voor kans-overeenstemming. Ondanks het feit dat de cijfers met 1 punt van elkaar mogen verschillen is de overeenstemming tussen beoordelaars zeer laag (T = 0.213). Voor de deelnemers aan de beoordelaarstraining kan geconcludeerd worden dat zij op basis van de kenmerk-beoordelingsmethode niet goed kunnen discrimineren tussen de kwaliteit van de aangeboden werkstukken.

Impressionistische (glance-and-grade) cijfers

Over de impressionistische cijfers worden dezelfde analyses uitgevoerd als over de op kenmerkbeoordelingen gebaseerde cijfers. Onderstaande tabel geeft een overzicht van de in de trainingssessies toegekende impressionistische cijfers.

werk	beoordelaar					werk	beoordelaar					werk	beoordelaar				
stuk	1	2	3	4	5	stuk	1	2	3	4	5	stuk	1	2	3	4	5
4	6	5	5	2	5	429	5	4	6	2	4	779	4	4	4	2	5
36	6	6	6	6	5	449	5	5	3	4	5	785	4	5	4	6	3
45	3	4	3	3	3		6	5	3	4	4	806	6	7	6	7	7
128	5	7	6	6	4	485	5	5	4	5	5	819	6	4	2	2	4
	6	7	4	6	5	498	6	5	4	5	4	868	5	7	6	4	5
202	4	5	3	4	2	537	6	6	5	4	6	869	4	6	5	5	6
229	5	4	2	4	3	541	6	5	6	5	7		6	6	5	4	7
257	5	6	7	6	6	644	6	7	6	6	6	893	5	5	4	5	6
278	4	5	3	5	4	658	6	5	6	6	6	897	4	6	5	6	6
288	4	4	2	3	3	728	5	6	3	4	4	931	6	5	4	8	6
305	5	6	5	7	6		3	5	3	3	4	954	7	7	7	6	7
	6	6	6	7	6	734	4	6	4	5	4		7	7	6	5	7
312	4	4	8	4	4	746	4	5	5	5	3	957	4	5	3	2	4
364	7	5	6	8	6		4	4	3	5	3	993	7	6	7	6	7
374	4	5	7	6	4	767	7	7	6	5	6	994	6	5	4	3	6
413	4	6	5	7	6		8	7	7	6	5						

De intraklasse correlatie coëfficiënt (ICC)

De tweeweg variantie analyse, nodig voor het berekenen van de ICC, levert de volgende variantie-schattingen op:

tweeweg variantie-analyse (mixed model)

bron	SS	df	MS	F
werkstukken	266.37	46	5.79	
beoordelaars	17.57	4	4.39	4.88*
error	165.63	184	0.90	
totaal	449.57	234		

* $p<.01$

De ICC is: $R = (5.79-0.90)/(5.79+(5-1)*0.90) = 0.52$. Deze R-waarde is significant op het 1 procent toetsings-niveau. De betrouwbaarheid van één beoordelaar is erg laag. De ICC voor de impressionistische cijfers is slechts weinig kleiner dan de ICC voor de cijfers die gebaseerd zijn op kenmerkbeoordelingen. Anders gezegd: het beoordelen van werkstukken aan de hand van de kenmerkmethode levert slechts een geringe winst op in termen van betrouwbare cijfers, ten opzichte van de glance-and-grade beoordeling.

index T

In de eerste plaats wordt gekeken of de nulhypothese - dat overeenstemming toevallig is - verworpen kan worden. Indien dit het geval is wordt de mate van overeenstemming uitgedrukt door middel van T. $\chi^2=534.016$, zodat de nulhypothese verworpen wordt. Index T bedraagt 0.189 en is dus weinig kleiner dan de overeenstemming die berekend werd over de cijfers gebaseerd op kenmerkscores.

Conclusie

Voor het geven van cijfers en met name voor het nemen van zak-/slaagbeslissingen zijn kenmerkbeoordelingen weinig betrouwbaarder dan impressionistische beoordelingen. De oorzaak hiervan moet gezocht worden in de onbetrouwbaarheid van de kenmerkbeoordelingen zelf. De criteria zijn te vaag omschreven om hoge overeenstemming tussen beoordelaars te kunnen bereiken. Zoals in par. 6.3.2 is aangetoond, wordt bij beoordeling op subkenmerken

een aanzienlijk hogere overeenstemming bereikt. Het is daarom aannemelijk dat cijfers gebaseerd op subkenmerkbeoordelingen betrouwbaarder zijn dan cijfers gebaseerd op kenmerkbeoordelingen. Helaas kan dit, als gevolg van de onvolledige waarnemingen op subkenmerkniveau, niet aangetoond worden met de beschikbare beoordelings-gegevens.

BIJLAGE 6

OVERZICHT VAN DE BEOORDELINGEN IN DE TRAININGS-SESSIES

OVERZICHT VAN DE BEOORDELINGEN UIT DE EERSTE TRAININGS-SESSIE

									OUTLINE								DIEPTE					CAV.OPP					CONVERG						P	AFWERKING							C*
Nr	As	Datum	O	D	C	C'	P	A	1	2	3	4	5	6	7	8	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	6	1	1	2	3	4	5	6	7	
288	00	820823	1	2	2	1	1	1	2	2	1	1	2	2	2	2	2	2	1	2	2	2	3	2	2	2	2	2	3	3	2	3	1	1	2	2	1	2	1	2	
288	1	821125	1	1	1	2	2	2									2	1	2	2	1	2	3	3	2	2	2	2	3	3	2	3	3	1	2	2	2	2	1	2	4
288	2	821125	1	2	1	1	2	2														2	3	3	3	2							3	1	2	2	2	2	2	2	4
288	3	821130	1	1	1	1	1	1									2	2	1	2	1	2	2	2	3	2															2
288	4	821207	1	3	1	1	2	1									2	2	1	2	2	2	2	1	1	2							2								3
288	5	821207	1	1	1	1	1	1									3	1	1	2	2	1	2	3	3	2															1
305	00	821112	1	3	3	3	1	2	2	2	2	2	2	1	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	3	1	2	2	2	2	2	2	
305	1	821125	2	3	2	2	1	2	2	2	2	2	2	1	1	2						2	2	2	2	2	2	2	2	3	2	2										5
305	2	821125	1	2	3	2	1	2									2	2	1	2	1						2	2	2	2	2	2									6
305	3	821130	1	3	2	2	1	2														2	2	2	2	2	2	2	2	3	2	2									6
305	4	821207	3	3	2	2	2	2	2	2	2	2	2	1	1	2						2	2	2	2	2	2	2	2	2	2	2	2									7
305	5	821207	1	3	2	2	2	2														2	2	3	2	2	2	1	2	2	2	2	3									6
728	00	820823	1	1	1	2	2	1	1	2	2	2	2	2	2	3	3	2	1	2	1	2	3	2	2	2	2	2	3	2	2	2	2	1	1	2	2	2	2	2	
728	1	821125	1	1	1	1	3	2														3	2	3	3	2	3	2	2	1	1	2	2	1	1	1	2	2	1	5
728	2	821125	1	1	3	1	3	2														1	3	3	1	3	3	2	3	3	1	3	1	2	2	2	2	2	2	6
728	3	821130	1	1	3	1	3	2														2	2	2	3	2	2	2	2	3	2	2	2	1	1	2	2	2	2	2	3
728	4	821207	1	2	1	1	2	1									3	2	1	2	1						3	2	3	1	2	3									4
728	5	821207	1	1	1	1	2	2																			3	2	3	2	2	3		2	2	2	2	1	2	2	4
734	00	821112	1	1	3	3	3	1	2	2	2	9	9	2	1	2	1	2	2	2	2	2	2	2	2	2	2	2	2	9	2	2	2	1	1	1	2	2	2	2	
734	1	821125	1	1	2	2	2	2														2	2	1	2	2	2	2	3	9	2	3	2	2	1	1	2	2	2	1	4
734	2	821125	1	2	2	1	1	2									3	2	1	2	2	2	2	2	2	2	2	2	3	9	2	2	2	1	2	2	2	2	2	2	6
734	3	821130	1	3	2	3	2	2									2	2	1	2	1	2	2	2	2	2							1	2	2	1	2	1	2	1	4
734	4	821207	2	2	2	1	2	2	2	2	1	9	9	2	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	9	2	2	2	2	2	2	3	2	2	2	5
734	5	821207	1	2	2	2	1	2									2	2	2	2	1	2	2	2	1	2	1	2	2	9	2	2	1	2	2	1	2	2	2	2	4
868	00	820823	2	3	2	3	2	2	2	2	2	9	9	2	2	2	2	2	2	2	2	3	3	1	2	2	2	2	2	9	2	2	1	2	2	2	2	2	2	2	
868	1	821125	1	3	2	2	3	2	2	2	2	9	9	2	1	2											2	2	2	9	2	3	2									5
868	2	821125	2	2	2	3	3	2									2	2	1	3	2												1								7
868	3	821130	1	3	2	3	2	2	2	2	2	2	9	2	1	2																										6
868	4	821207	2	3	2	2	2	2																			2	2	2	9	2	2									4
868	5	821207	2	1	3	3	2	1									2	9	2	3	2	2	2	2	2	2								2	1	1	2	3	2	1	5
869	00	820823	3	2	2	1	3	2	2	2	2	9	9	2	2	2	3	3	2	2	1	2	2	2	1	2	2	2	3	9	2	2	2	1	2	2	2	2	2	2	
869	1	821125	1	3	2	2	2	2	2	2	2	9	9	2	1	2	2	2	2	2	2						2	2	2	9	2	3	2									4
869	2	821125	2	3	2	3	1	1	1	2	2	9	9	2	2	2	2	2	2	1	1						1	2	2	9	2	3	2	1	2	2	2	2	2	2	6
869	3	821130	1	2	2	2	2	2	2	2	2	9	9	1	1	1											2	2	2	9	2	3	3									5
869	4	821207	2	1	2	2	3	2	2	2	2	9	9	2	1	2	2	2	1	2	2						2	2	1	9	2	2									5
869	5	821207	2	2	3	2	1	2	2	2	2	9	9	1	1	1						2	2	2	2	2	2	3	2	9	2	2	1									6

Nr = werkstuknummer;
As = assistentnummer (00 = referentie-oordeel);
O = kenmerk outline; C' = kenmerk convergentie;
D = kenmerk diepte; P = kenmerk/subkenmerk
C_o = kenmerk cav. opp. hoek; pulpo a. afschuining;
C* = impres. totaaloordeel; A = kenmerk afwerking.

OVERZICHT VAN DE BEOORDELINGEN UIT DE TWEEDE TRAININGS-SESSIE

									OUTLINE							DIEPTE					CAV.OPP					CONVERG						P	AFWERKING							C*
Nr	As	Datum	O	D	C	C'	P	A	1	2	3	4	5	6	7	8	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	6	1	1	2	3	4	5	6	7	
36	00	821112	1	1	3	3	1	3	2	2	2	9	9	3	2	2	2	2	2	1	2	2	2	2	2	2	2	2	2	9	2	2	1	3	3	3	3	3	3	3	
36	1	830103	1	3	3	2	2	1									2	2	2	1	1						2	2	3	9	2	2	2	3	2	1	1	2	3	2	6
36	2	830105	2	2	2	3	1	1	2	3	2	9	9	2	1	2	2	2	2	2	1	2	2	2	2	2								1	2	1	1	2	1	1	6
36	3	830104	1	2	3	3	2	2									3	3	2	1	2												1	3	2	1	1	3	2	2	6
36	4	830104	2	3	3	2	3	2	2	2	2	9	9	3	1	2	2	2	1	1	2						2	2	1	9	2	1	2	1	3	1	1	3	2	2	6
36	5	830111	2	2	2	1	1	2	2	2	2	9	9	3	1	2	2	2	2	1	2	3	2	2	3	2	2	3	2	9	2	2		2	3	1	1	3	2	2	5
429	00	820823	1	1	1	1	2	2	1	2	1	2	9	2	3	1	2	1	2	2	1	3	3	2	2	1	3	2	2	2	3	1	2	2	2	2	2	2	1	2	
429	1	830103	2	1	2	1	2	2	2	2	1	2	9	3	3	1						2	2	2	2	3																5
429	2	830105	1	1	2	1	1	2														2	2	2	2	2							2								4
429	3	830104	1	1	1	1	1	1																									1	1	1	2	2	2	2	2	1
429	4	830104	1	1	2	1	2	2														3	3	3	1	2															2
429	5	830111	1	1	3	1	2	2														2	3	2	2	2															4
658	00	821112	2	2	3	2	3	2	2	2	2	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	3	1	3	2	1	2	1	2	2	2	2	
658	1	830103	1	2	2	2	3	2	2	2	1	1	1	1	1	2						2	2	2	2	2															6
658	2	830105	1	2	3	2	2	2	2	2	1	1	1	2	2	2																	2								5
658	3	830104	1	3	3	2	2	1	2	2	1	1	1	1	2	2	2	2	2	2	1												2	1	3	1	2	2	2	2	6
658	4	830104	3	3	3	2	3	2	2	2	1	2	1	2	2	2	2	2	1	1	2																				6
658	5	830111	1	3	3	2	2	2	2	2	1	1	2	1	1	1	2	2	2	2	1												1								6
785	00	821112	2	2	3	1	1	2	2	2	2	2	2	1	2	2	3	2	1	2	2	2	2	2	2	2	2	2	3	3	3	3	1	2	2	2	1	2	2	2	
785	1	830103	1	1	3	1	2	3	2	2	1	1	9	1	1	2	3	3	1	3	2												1	3	3	2	3	2	2	3	4
785	2	830105	1	3	2	1	1	3	2	3	2	1	9	1	2	2	2	2	2	2	2	2	2	2	2	2								1	3	2	1	1	1	3	5
785	3	830104	1	1	3	1	1	2	2	3	2	1	9	1	2	2	3	3	2	3	1																				4
785	4	830104	2	1	3	1	1	3									3	3	1	2	2													3	3	3	3	2	2	3	6
785	5	830111	1	1	2	1	1	2	2	3	2	2	2	1	1	2	3	3	2	2	1	2	2	2	3	2															3
819	00	820823	1	1	2	2	2	3	2	3	1	2	2	2	1	2	2	1	2	1	2	2	3	2	2	2	2	2	3	3	2	2	1	3	3	3	3	3	3	2	
819	1	830103	2	1	2	2	1	3	2	2	1	2	9	2	2	2																	3								6
819	2	830105	1	1	2	1	1	2																			3	3	3	3	2	2	2	1	2	2	1	2	2	1	4
819	3	830104	1	1	1	1	1	2														2	2	2	3	2	3	3	3	3	2	1	1	1	2	2	2	2	2	2	2
819	4	830104	1	1	2	1	2	2																			3	3	2	3	2	2		2	2	1	1	2	2	1	2
819	5	830111	1	1	2	1	1	3																			2	3	3	2	2	2	1								4
869	00	820823	3	2	2	1	3	2	2	2	2	9	9	2	2	2	3	3	2	2	1	2	2	2	1	2	2	2	3	9	2	2	2	1	2	2	2	2	2	2	
869	1	830103	1	3	1	2	3	2	2	2	2	9	9	2	1	2	2	2	2	2	2	2	2	1	1	2	2	2	2	9	2	3									6
869	2	830105	2	3	2	2	2	1	2	2	2	9	9	2	1	2	2	2	2	2	2						2	2	2	9	2	3	2	1	2	1	1	1	1	2	6
869	3	830104	2	2	2	2	2	2	2	2	2	9	9	2	1	1											2	2	2	9	2	3	2								5
869	4	830104	2	3	2	1	3	2	2	2	2	9	9	2	1	2	2	2	1	2	2																				4
869	5	830111	1	3	2	3	3	2	2	2	2	9	9	1	1	1	2	2	1	2	2						2	2	2	9	2	2									7
954	00	821112	3	2	3	2	3	2	2	2	2	9	9	2	2	2	3	2	2	2	2	2	2	2	2	2	2	2	2	9	_	2	2	2	2	1	2	2	2	2	
954	1	830103	2	3	2	3	3	1	2	2	2	9	9	2	1	2	3	2	2	2	2	2	2	2	1	2	2	2	2	9	2	3		2	2	1	2	2	2	2	7
954	2	830105	2	3	3	3	3	1	2	2	2	9	9	2	2	2	3	2	2	2	2						2	2	2	9	2	2		2	2	1	1	2	2	2	7
954	3	830104	2	2	2	3	3	2	2	2	2	9	9	2	1	2						2	2	2	1	2	2	2	1	9	2	2									7
954	4	830104	2	3	2	2	3	2	2	2	3	9	9	2	1	2	2	2	2	2	2	2	1	1	2	2															6
954	5	830111	2	3	2	3	3	2	2	2	2	9	9	2	1	2	2	2	2	3	2	2	2	2	1	2	2	2	2	9	2	2									7

OVERZICHT VAN DE BEOORDELINGEN UIT DE PERDE TRAINING-SESSIE

									OUTLINE							DIEPTE					CAV.OPP					CONVEXC						P	AFJERKING							
Nr	As	Datum	O	D	C	C	P	A	1	2	3	4	5	6	7	8	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	6	1	1	2	3	4	5	6	7	C*
004	00	820823	3	2	1	1	1	2	2	2	2	9	9	2	2	2	3	3	2	3	1	2	2	2	1	2	2	2	3	9	2	3	1	1	2	2	2	2	2	2	
004	1	830124	2	1	1	2	2	2	2	2	2	9	9	2	2	3	2	3	2	3	1						2	2	3	9	2	3	1								6
004	2	830124	2	1	2	2	1	2	2	2	2	9	9	2	2	2	3	3	2	3	1	2	2	2	1	2	2	2	2	9	2	3									5
004	3	830125	2	1	1	2	2	2	1	2	2	9	1	2	2	3	3	3	2	3	2						2	2	3	9	2	2	1								5
004	4	830121	1	3	1	2	1	2	2	2	2	9	9	2	2	2	2	2	2	3	1						2	2	2	9	2	3									2
004	5	830125	2	1	3	2	1	2	2	2	2	9	9	2	2	2	2	3	2	3	1	2	2	2	2	2	1	2	2	9	2	2									5
498	00	820823	2	3	2	1	3	2	2	2	1	2	9	1	2	2	2	2	2	2	2	2	3	2	2	2	2	2	3	3	2	1	2	1	2	2	2	2	2		
498	1	830124	2	2	2	1	2	1									2	2	2	2	2												2	2	2	1	2	2	2	2	6
498	2	830124	2	2	1	1	2	2									2	2	2	2	2	2	3	2	2	2							2								5
498	3	830125	2	1	2	1	3	1									2	2	2	2	2													1	1	1	2	2	2	2	4
498	4	830121	1	2	2	1	3	2	2	2	1	2	9	2	1	2	1	2	1	2	2																				5
498	5	830125	1	1	1	1	2	1	2	2	1	2	9	2	1	2	2	2	1	2	2	2	3	2	2	2							2	2	2	1	2	1	1	2	4
537	00	830114	3	2	2	1	3	2	2	2	2	9	9	2	2	2	3	3	2	2	1	2	2	2	1	2	2	2	3	9	2	2	2	1	2	2	2	2	2		
537	1	830124	2	2	1	2	3	2	2	2	2	9	9	2	1	2	2	2	2	1	2	2	2	1	2	2	2	2	2	9	2	3									6
537	2	830124	2	2	3	2	2	2	2	2	2	9	9	2	1	2	2	2	2	2	2						2	2	2	9	2	3	3								6
537	3	830125	2	2	1	2	3	2	2	2	2	9	9	2	1	1						2	2	2	1	2	2	2	1	9	2	3									5
537	4	830121	1	3	2	2	2	2	2	2	2	9	9	1	1	1	2	2	2	2	1						2	2	2	9	2	3	1								4
537	5	830125	1	2	2	3	2	2	2	2	2	9	9	1	1	2											2	2	2	9	2	3	3								6
746	00	821112	1	1	1	1	1	3	2	2	2	2	9	1	1	2	3	2	3	3	2	2	2	3	3	2	2	2	2	3	2	2	1	3	3	3	3	3	3	3	
746	1	830124	1	1	1	1	1	2																										2	2	2	2	2	1	2	4
746	2	830124	1	2	2	2	1	1									3	3	2	2	2	2	2	3	3	2	2	2	3	3	3	2		1	3	1	1	1	2	1	5
746	3	830125	1	2	2	1	2	2									3	3	2	2	2	2	2	2	2	2							1	2	2	1	2	1	2	2	5
746	4	830121	1	3	1	2	2	2									2	2	2	2	2						2	2	2	3	2	2	1	3	2	3	3	1	2	2	5
746	5	830125	1	1	1	1	1	2																										1	2	2	3	1	1	2	3
767	00	821112	3	2	3	3	3	2	2	2	2	9	9	1	1	1	3	2	2	2	2	2	2	2	2	2	2	2	2	9	2	2	2	2	2	1	2	2	2	2	
767	1	830124	3	2	3	2	3	2																			2	2	2	9	2	2									7
767	2	830124	2	2	3	3	3	2	2	2	2	9	9	1	1	1																										7
767	3	830125	2	1	2	3	3	1	2	2	2	9	9	1	1	1	3	2	2	2	2	2	2	2	2	2								2	3	1	2	3	2	2	6
767	4	830121	1	3	2	2	3	2	2	2	2	9	9	1	1	1	2	2	2	2	2	2	2	2	2	2	2	2	1	9	2	2									5
767	5	830125	1	1	2	3	3	2	2	2	2	9	9	1	1	1	3	2	2	2	2	2	2	2	2	2															6
779	00	821112	1	1	2	1	2	3	2	3	1	2	9	2	2	2	1	1	1	2	2	2	2	2	3	2	2	2	3	3	2	3	1	3	3	3	3	3	3	3	
779	1	830124	1	1	2	1	2	2																										2	2	2	2	3	2	2	4
779	2	830124	1	1	2	1	1	2																									2	1	2	2	1	1	2	2	4
779	3	830125	1	1	2	1	2	2																										1	2	2	2	2	2	2	4
779	4	830121	1	2	1	1	1	2									2	1	1	2	1	2	2	2	3	2							1	2	1	1	2	2	1	2	2
779	5	830125	1	1	2	2	2	2																			2	2	2	3	2	3		2	2	2	2	2	1	1	5
954	00	821112	3	2	3	2	3	2	2	2	2	9	9	2	2	2	3	2	2	2	2	2	2	2	2	2	2	2	2	9	2	2	2	2	2	1	2	2	2	2	
954	1	830124	2	2	3	3	3	2	2	2	2	9	9	2	1	2											2	2	2	9	2	3									7
954	2	830124	2	3	3	3	3	2	2	2	2	9	9	2	1	2	2	2	2	2	2						2	2	2	9	2	2									7
954	3	830125	2	1	1	3	3	2	2	2	2	9	9	2	1	1	3	2	2	2	2	2	2	2	2	2	2	2	2	9	2	2									6
954	4	830121	1	3	2	2	3	1	2	2	2	9	9	2	1	2	2	2	2	2	2	2	2	3	2	2								3	2	1	2	2	2	2	5
954	5	830125	2	3	2	3	2	3	2	2	2	9	9	2	1	2	2	2	2	2	2	2	2	2	1	2	2	2	2	9	2	2	2	3	3	1	3	2	2	2	7
957	00	820823	1	2	3	1	3	2	2	2	1	2	2	2	2	1	2	2	1	2	1	2	2	2	2	2	3	2	3	3	2	3	2	1	1	2	2	2	2	2	
957	1	830124	1	1	2	1	2	2									2	1	1	2	2	2	3	2	1	2							2								4
957	2	830124	1	2	2	1	2	2														2	3	2	2	2							2								5
957	3	830125	1	1	2	1	2	1									3	2	2	2	2	2	2	2	2	2							2	1	1	1	2	1	2	1	3
957	4	830121	1	2	2	1	1	1														2	2	1	2	3							2	1	1	1	1	1	2	2	2
957	5	830125	1	1	2	1	2	2									3	2	1	2	1	2	3	2	2	2							3								4

OVERZICHT VAN DE RESULTATEN UIT DE VIERDE TRAININGS-SESSIE

Nr	As	Datum	O	D	C	C'	P	A	OUTLINE 1	2	3	4	5	6	7	8	DIFFER 1	2	3	4	5	CAV.OPP 1	2	3	4	5	CONVERG 1	2	3	4	5	6	P	AANMERKING 1	2	3	4	5	6	7	C*
128	00	821112	1	3	2	3	1	2	2	2	2	9	9	1	1	2	2	2	2	2	2	2	2	2	1	2	?	?	?	?	?	?	1	2	2	1	2	2	2	2	
128	1	830203	1	3	1	3	2	1														?	2	2	1	2							1	3	2	*	2	3	2	2	5
128	2	830202	2	3	2	3	3	2	2	2	2	9	9	1	1	1																	1								7
128	3	830210	2	3	2	3	1	2	2	2	2	9	9	1	1	1																										6
128	4	830207	1	3	2	2	2	2																		?	3	1	9	3	2	1									6
128	5	830208	1	2	1	2	1	2									?	2	2	2	1	2	?	?	1	2	3	2	2	9	?	?										4
202	00	821112	1	1	1	1	1	1	1	2	1	2	?	?	?	?	2	2	1	?	?	2	?	3	2	?	?	?	?	3	3	2	?	3	?	1	1	?	2	2	1	
202	1	830203	1	1	1	2	1	1																									3	?	?	3	3	?	3		4
202	2	830202	1	3	2	1	2	1									?	?	2	?	?	?	?	3	2	?	?							3								5
202	3	830210	1	2	2	1	1	1									2	?	2	?	?	?	?	3	2	?	?															3
202	4	830207	1	3	1	1	1	1									?	?	1	?	?																					4
202	5	830208	1	1	1	1	1	1																																		?
229	00	820823	1	1	2	2	2	1	1	3	?	?	1	?	?	?	1	1	1	?	?	?	3	?	2	?	3	?	?	2	?	1	2	?	1	1	1	?	?	?	?	
229	1	830203	3	1	1	1	2	2	2	?	?	1	9	3	2	?						2	3	?	1	?	?	?	3	?	2	3		1	1	?	1	?	?	?	5
229	2	830202	1	1	2	1	2	2																			?	3	3	?	?	3		1	1	1	1	1	1	1	4
229	3	830210	1	1	2	1	1	1																			?	?	3	?	2	3	?									?
229	4	830207	1	1	2	1	2	1																			?	?	3	?	?	3										4
229	5	830208	1	1	2	1	1	2																			3	2	3	2	?	3	1	1	1	?	1	?	?	?	3
257	00	830202	3	2	2	1	3	2	?	2	2	9	9	?	?	2	3	3	2	2	1	?	?	?	1	2	?	?	3	9	?	?	2	1	?	?	?	?	?	?	?	
257	1	830203	1	3	1	1	1	2	2	2	2	9	9	2	1	2	3	3	?	?	?	2	2	2	1	1	?															5
257	2	830202	2	3	2	?	3	2	2	2	?	9	9	2	1	?	?	2	?	?	2	2						?	2	2	9	2	?										6
257	3	830210	?	3	2	3	3	?	?	?	2	9	9	2	1	1	2	3	2	2	2							2	3	?	9	?	?										7
257	4	830207	1	3	1	2	2	2	?	2	2	9	9	1	1	1	2	?	1	?	1	?	2	2	1	?	?	2	?	1	9	?	?	2									6
257	5	830208	2	1	2	2	1	2	?	2	2	9	9	2	1	?	3	?	?	?	?							2	2	2	9	?	?										6
278	00	820823	1	1	3	3	2	2	?	2	2	2	2	1	2	3	?	?	?	1	1	?	?	?	?	?	2	?	?	?	?	?	?	1	2	2	1	?	?	?	1	
278	1	830203	1	1	2	3	2	1														?	2	?	2	2									2	?	?	1	2	1	1	4
278	2	830202	1	1	2	2	1	2														2	2	?	?	?	?	?	?	2	2	2	1									5
278	3	830210	1	1	2	?	1	?														3	2	2	?	?	2	2	?	?	?	?	?	1								3
278	4	830207	2	2	1	3	1	1	2	2	?	?	3	1	1	?	1	1	?	1	1	2	?	3	?	?							1	2	2	1	1	1	1	1	5
278	5	830208	1	1	2	2	1	?														?	?	3	?	?	?	?	?	3	?	2	1									4
767	00	821112	3	?	3	3	3	2	2	2	?	9	9	1	1	1	3	?	?	?	?	2	?	?	?	2	2	?	?	9	?	?	?	?	?	?	1	2	?	?	2	
767	1	830203	2	3	2	3	3	2	?	?	2	9	9	2	2	2	3	3	2	3	2	?	?	?	2	2																8
767	2	830202	2	2	3	3	3	3	?	?	2	9	9	1	1	1																					2	?	1	1	2	1	1	7
767	3	830210	3	1	3	3	3	3									3	?	?	?	?	2															2	?	1	1	?	?	2	7
767	4	830207	1	3	2	3	3	2	2	?	?	9	9	1	1	1	?	?	?	2	2	2	?	3	?	?																6
767	5	830208	1	1	3	3	3	2	2	?	3	9	9	1	1	1	3	2	?	3	?																					5
993	00	820823	3	2	1	3	3	3	2	?	2	9	9	1	1	2	?	?	?	?	1	2	2	?	1	?	?	?	2	9	2	2	?	3	3	3	3	3	3	3	
993	1	830203	2	3	2	2	3	3	2	2	2	9	9	?	?	1	2	?	?	?	?	?	?	2	1	1	2	?	?	?	9	2	1										7
993	2	830202	2	3	3	2	3	3	2	?	?	9	9	2	?	?	?	?	?	?	2	2	?	?	1	?	2	?	2	2	9	?	?										6
993	3	830210	3	2	2	3	2	3														?	?	1	1	?							?									7
993	4	830207	1	3	2	2	?	?	?	3	2	9	9	1	1	1	2	?	?	?	?	?	?	3	?	?	?	2	1	9	?	?	2	2	2	1	?	3	2	3	6
993	5	830208	2	2	3	?	3	3	?	?	3	9	9	?	1	?						2	?	?	?	?	3	?	?	9	?	?										7
994	00	821112	1	3	3	2	2	2	?	3	2	2	2	?	?	?	?	?	?	2	2	?	?	?	?	2	?	?	?	?	?	?	3	?	?	1	2	2	2	2	
994	1	830203	2	3	?	?	2	1	2	3	1	?	9	3	2	?						2	2	2	2	?									2	2	2	?	2	3	1	6
994	2	830202	1	3	2	1	2	2														2	3	2	?	?	2	2	3	2	?	?										5
994	3	830210	1	?	3	1	1	2									?	2	2	2	?						2	2	3	2	?	?	2									4
994	4	830207	1	3	2	1	1	1														?	?	?	3	3	?	?	3	2	?	3		?	?	1	1	2	1	1	3
994	5	830208	1	2	2	2	2	2									2	?	?	?	1	2	3	2	?	2																6

OVERZICHT VAN DE BEOORDELINGEN UIT DE VIJFDE TRAININGS-SESSIE

Nr	As	Datum	O	D	C	C'	P	A	OUTLINE								DIEPTE					CAV.OPP					CONVERG						P	APPERTING						C*
									1	2	3	4	5	6	7	8	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	6	1	1	2	3	4	5	6	7	
45	00	821112	1	1	1	1	1	2	2	2	2	9	9	1	2	2	3	3	2	3	2	2	2	3	2	2	2	2	3	9	2	1	1	2	2	1	2	2	2	2	
45	1	830223	1	1	1	2	1	1																			2	2	2	9	2	1		3	2	1	2	2	2	1	3
45	2	830225	2	1	1	2	1	2	1	2	2	9	9	1	2	3											2	2	2	9	2	2									4
45	3	830224	2	1	1	1	1	2	1	2	2	9	9	1	1	1																									3
45	4	830228	1	1	1	1	2	1																									2	2	1	1	1	1	1	1	3
45	5	830301	1	1	1	2	1	2																			3	2	2	9	2	2									3
128	00	821112	1	3	2	3	1	2	2	2	2	9	9	1	1	2	2	2	2	2	2	2	2	2	1	2	2	2	2	9	2	2	1	2	2	1	2	2	2	2	
128	1	830223	2	2	1	3	2	2	2	2	2	9	9	1	1	1	2	2	2	2	2	2	2	2	1	2						1									6
128	2	830225	3	3	3	3	1	2	2	2	2	9	9	1	1	1						2	2	2	2	2															7
128	3	830224	1	3	1	2	1	2														2	2	2	1	2	2	2	2	9	2	2									4
128	4	830228	1	2	2	3	3	2									2	2	3	2	2												2								6
128	5	830301	1	2	2	2	1	1									2	2	2	2	1						2	2	2	9	2	2		2	2	1	1	3	2	2	5
374	00	820823	1	3	1	1	1	2	2	3	1	2	9	1	1	2	2	2	2	2	2	3	2	1	3	2	2	3	2	2	3	2	3	2	2	2	2	2	2	2	
374	1	830223	1	2	1	1	3	2									2	2	2	2	2												2								4
374	2	830225	1	2	2	1	3	2									2	3	2	2	2	1	2	2	3	2							2								5
374	3	830224	1	3	1	1	3	2																									1								5
374	4	830228	1	2	1	1	2	2									2	2	2	1	2												2								6
374	5	830301	1	1	1	1	2	2									2	2	1	2	2												2								4
413	00	830214	3	2	2	1	3	2	2	2	2	9	9	2	2	2	3	3	2	2	1	2	2	2	1	2	2	2	3	9	2	2	2	1	2	2	2	2	2	2	
413	1	830223	1	1	2	1	2	1	2	2	2	9	9	2	1	2	2	2	2	2	1												2	2	2	2	2	2	2	2	4
413	2	830225	3	3	2	2	2	2									2	2	2	2	2						2	2	2	9	2	2	2								6
413	3	830224	2	1	1	3	3	3	2	2	2	9	9	2	1	2	3	2	2	2	2	2	2	2	1	2	2	3	2	9	2	3		2	2	1	2	2	2	2	6
413	4	830228	1	3	2	3	2	2	2	2	2	9	9	1	1	1	2	2	2	2	2						2	2	2	9	2	2	2									7
413	5	830301	1	2	2	2	2	2	2	2	2	9	9	2	1	2											2	2	3	9	2	2	2									6
449	00	821112	1	2	3	2	2	2	2	2	1	2	9	1	2	2	2	2	1	2	2	2	2	2	2	2	2	2	2	3	2	2	2	2	2	2	2	2	2	2	
449	1	830223	1	2	2	1	3	2														2	2	2	2	2	3	2	2	3	2	2	2								5
449	2	830225	1	1	2	2	2	2									2	2	1	2	2	2	2	2	2	2															5
449	3	830224	1	1	2	1	3	1									3	2	1	2	1	2	2	2	2	2	3	2	2	3	2	2	2	1	2	1	2	1	1	2	3
449	4	830228	1	3	2	1	3	1									2	2	2	2	2	2	2	2	3	2	3	2	2	3	2	2	2	2	1	2	1	1	2	4
449	5	830301	1	2	2	1	3	2														2	2	3	2	2	3	2	2	3	2	2	2								5
485	00	821112	1	3	3	3	3	3	2	2	2	2	2	1	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	3	3	3	3	3	3	3	
485	1	830223	1	3	2	2	1	2														2	2	2	2	2	2	2	2	2	2	3	3	3	3	2	3	2	2	5
485	2	830225	2	2	2	2	2	2	2	2	2	2	1	1	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	5
485	3	830224	1	2	2	1	2	2									3	2	2	2	2	2	2	2	2	2	2	2	3	2	2	2		1	2	1	2	2	2	2	4
485	4	830228	1	3	2	2	2	2														2	2	3	2	2	3	3	2	2	2	2	2	2	2	1	3	2	2	5
485	5	830301	1	2	2	1	2	2									2	2	2	2	1	2	2	3	2	2	3	2	2	2	2	2	2	2	2	1	2	2	2	3	5
728	00	820823	1	1	1	2	2	1	1	2	2	2	2	2	2	3	3	2	1	2	1	2	3	2	2	2	2	2	3	2	2	2	2	2	1	1	2	2	2	2	
728	1	830223	1	1	1	1	2	2																			2	2	3	2	2	3		3	2	2	2	1	2	2	3
728	2	830225	1	2	2	1	2	2									3	3	2	2	2	2	1	2	2	2	2	1	3	2	1	3		1	1	1	1	1	1	1	5
728	3	830274	1	1	1	1	3	2																			2	2	3	1	2	3	2	2	1	1	1	2	2	2	3
728	4	830228	1	2	1	1	2	1									2	2	1	2	2						3	2	3	1	2	3									3
728	5	830301	1	1	1	1	2	2																			3	2	3	1	2	2		2	1	2	2	2	2	2	4
806	00	820823	1	1	3	3	2	3	2	3	2	9	9	2	1	2	2	2	1	2	1	2	2	2	2	2	2	2	2	9	2	2	2	2	2	2	2	2	2	2	
806	1	830223	2	2	1	2	3	2	2	2	2	9	9	2	1	2	2	2	2	2	1	2	2	2	1	2	2	2	2	9	2	2	2	2	2	2	2	2	2	2	6
806	2	830225	2	3	3	3	3	2	2	2	2	9	9	2	1	2	2	2	2	2	2												2	2	2	1	2	2	2	2	7
806	3	830224	2	1	2	3	3	3	1	2	2	9	9	3	1	2						2	2	2	2	2							2								6
806	4	830228	2	1	2	3	3	2	2	2	3	9	9	2	1	1						2	1	3	2	2							2	2	2	1	2	2	2	2	7
806	5	830301	1	2	2	3	3	3									2	2	2	2	1	2	2	2	3	2							2								7
931	00	820823	2	2	2	3	2	2	2	2	2	2	2	2	1	2	2	2	1	2	2	3	3	2	2	2	2	2	2	2	2	2	1	2	2	2	2	2	2	2	
931	1	830223	2	1	2	2	2	2									2	1	2	2	2						2	2	2	3	2	2									6
931	2	830225	1	2	2	2	2	2	1	2	2	1	1	2	2	2											2	3	2	2	2	2									5
931	3	830224	2	1	3	1	2	1									2	1	2	2	2	2	2	2	2	2	2	2	3	2	2	2		2	2	1	2	1	2	2	4
931	4	830228	1	3	2	3	2	2	2	2	2	2	1	2	1	2	2	2	2	2	2																				8
931	5	830301	2	2	3	2	2	2														2	2	2	2	2	2	2	2	3	2	2									6

OVERZICHT VAN DE BEOORDELINGEN UIT DE ZESDE TRAININGS-SESSIE

									OUTLINE								DIPPIE					CAV.OPP					CONVEXO						P	AFWERKING							
Nr	As	Datum	O	D	C	C'	P	A	1	2	3	4	5	6	7	8	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	6	1	1	2	3	4	5	6	7	C*
305	00	821112	1	3	3	3	1	2	2	2	2	2	2	1	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	3	1	2	2	2	2	2	2	
305	1	830307	1	2	2	2	2	2									2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	3								6
305	2	830307	2	3	2	3	2	2	2	1	2	2	2	2	2	2						2	2	2	2	2							3								6
305	3	830310	3	3	3	2	2	3	2	2	2	2	2	1	1												2	2	3	2	2	2	3	1	2	2	2	2	2	2	8
305	4	830315	2	1	3	3	1	3	2	2	2	2	2	1	1	2																			3	2	2	3	2	2	2	7
305	5	830322	1	2	2	2	3	3									2	2	2	2	1	2	2	3	2	2	2	2	2	3	2	2	2	2	2	2	2	2	2	2	6
312	00	820823	1	1	1	2	1	2	1	2	2	1	1	2	2	2	3	3	1	2	1	3	3	1	2	2	3	2	2	2	2	2	1	2	2	2	2	2	2	2	
312	1	830307	1	1	2	2	1	2														2	3	3	2	2															4
312	2	830307	1	1	2	2	1	3														2	2	2	1	2								2	2	2	1	2	2	2	4
312	3	830310	1	1	3	1	2	3														2	3	2	2	2	2	3	3	3	2	2	2	1	2	2	1	2	2	2	6
312	4	830315	1	3	2	2	3	2									2	2	2	2	2	2	2	2	3	2							1	1								4
312	5	830322	1	1	2	1	1	2														2	2	2	3	2	3	2	2	2	2	2									4
364	00	821112	2	3	3	2	3	3	2	2	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	3	3	3	3	3	3	3	
364	1	830307	2	2	2	2	2	2									2	2	2	2	2	2	2	2	2	2							2	2	2	2	2	2	2	2	7
364	2	830307	1	2	2	2	2	2	1	2	1	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2							2	1	2	2	2	2	1	2	5
364	3	830310	2	3	3	3	3	2																			2	2	2	3	2	2		1	2	1	2	2	2	2	7
364	4	830315	3	3	3	3	3	2	2	2	2	3	2	1	2	2											2	2	2	3	2	2		2	2	2	2	2	2	3	8
364	5	830322	2	3	2	2	2	2														2	2	3	2	2							3	2	2	1	2	2	3	2	6
449	00	821112	1	2	3	2	2	2	2	2	1	2	9	1	2	2	2	2	1	2	2	2	2	2	2	2	2	2	2	3	2	2	2	2	2	2	2	2	2	2	
449	1	830307	1	2	3	2	2	2																																	6
449	2	830307	1	3	3	3	3	2									2	2	2	2	2						2	2	2	2	2	2	2								5
449	3	830310	1	1	3	1	3	1									3	2	1	2	2						3	2	2	3	2	2	2	2	2	2	1	2	2	3	3
449	4	830315	1	2	3	1	3	1																			3	2	2	3	2	2	2	2	2	1	1	2	2	4	
449	5	830322	1	2	2	1	3	1														2	2	2	3	2	2	2	2	3	2	2	2	2	2	2	1	2	2	2	4
541	00	820823	1	1	3	3	2	2	2	3	2	1	9	3	2	2	1	1	2	3	2	2	2	2	2	2	2	2	2	2	2	2	1	3	2	2	2	3	3	1	
541	1	830307	3	1	2	2	2	2	2	2	1	2	9	3	2	2						2	2	2	2	2	2	2	3	3	2	2										6
541	2	830307	1	2	2	1	2	2									2	2	2	2	1	2	2	2	2	2	2	2	3	2	2	3									5
541	3	830310	1	3	2	1	3	3									2	1	1	2	2	2	2	2	2	2	2	2	3	2	2	2	2	2	1	2	2	2	2	2	6
541	4	830315	2	2	2	2	2	2	2	3	1	2	9	1	2	2	1	1	2	2	2	2	2	2	2	2	3	2	2	3	2	2									5
541	5	830322	2	2	2	3	3	2	2	2	2	2	2	2	2	2	2	2	2	2	2																				7
644	00	830304	3	2	2	1	3	2	2	2	2	9	9	2	2	2	3	3	2	2	1	2	2	2	1	2	2	2	3	9	2	2	2	1	2	2	2	2	2	2	
644	1	830307	2	3	2	2	2	2	2	2	2	9	9	2	1	2	2	2	2	2	2						2	2	2	9	2	3	2									6
644	2	830307	2	3	3	3	2	2	2	2	2	9	9	2	1	2	2	2	2	2	1	2	2	2	2	2	2	2	2	9	2	2	2								7
644	3	830310	2	2	2	3	1	2	2	2	2	9	9	2	1	1											2	2	3	9	2	2									6
644	4	830315	2	3	2	3	2	3	2	2	2	9	9	1	1	1	2	2	1	2	2						2	2	2	9	2	2	3	3	2	2	2	2	3	2	6
644	5	830322	1	2	2	2	2	2	2	2	2	9	9	1	1	1											2	2	2	9	2	2	3								6
746	00	821112	1	1	1	1	1	3	2	2	2	2	9	1	1	2	3	3	2	3	2	2	2	3	3	2	2	2	2	3	2	2	1	3	3	3	3	3	3	3	
746	1	830307	1	1	1	2	1	2																			2	2	2	3	3	2		2	2	2	2	1	1	2	4
746	2	830307	1	2	2	2	1	3									3	3	2	2	2	2	2	3	2	2	2	2	2	2	3	2										4
746	3	830310	1	1	1	2	1	1																			2	2	3	2	3	1		2	2	2	2	1	2	1	1
746	4	830315	1	2	1	2	1	2									3	3	2	3	2	2	2	2	3	2	2	2	3	2	2			2	3	3	2	3	3	2	5
746	5	830322	1	1	1	2	1	2																			2	2	2	3	2	2		2	1	2	2	2	2	2	3
893	00	821112	1	3	3	3	3	2	2	2	1	9	9	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	9	2	2	2	2	2	1	2	2	2	2	
893	1	830307	1	2	3	3	3	2									2	2	2	2	2																				5
893	2	830307	1	1	3	3	3	2									2	2	2	3	2																				5
893	3	830310	1	1	3	1	3	2									2	2	2	3	2						2	2	3	9	2	2									4
893	4	830315	2	3	3	3	2	2	2	2	1	9	9	2	1	1																	1								5
893	5	830322	2	1	3	3	2	3	2	2	1	9	9	2	2	2	3	3	2	2	2												1	3	2	1	2	3	2	2	6
897	00	820823	1	2	2	1	1	2	2	2	2	9	9	2	1	1	2	2	2	2	1	2	1	2	2	2	2	2	3	9	2	2	3	1	2	2	2	2	2	2	
897	1	830307	1	2	2	1	3	1																									2	2	2	1	1	2	2	2	4
897	2	830307	2	2	2	2	3	2	2	2	2	9	9	2	1	2											2	2	1	9	2	3	2								6
897	3	830310	2	2	2	1	2	2	2	2	2	9	9	2	1	2																	3								5
897	4	830315	2	2	2	3	3	1	2	2	2	9	9	1	1	1											2	2	2	9	2	2	2	2	2	1	1	2	2	1	6
897	5	830322	1	2	2	3	3	2																			2	2	2	9	2	2	2								6

BIJLAGE 7

VRAGENLIJST TER INVENTARISATIE VAN DE MENINGEN VAN STUDENTEN OVER PMP'S EN PPP'S

management probleem:
student nummer:

Beantwoord de onderstaande vragen door het aankruisen van het cijfer dat uw mening het beste weergeeft.

---

1. In welke mate heeft het PMP/PPP uw motivatie geprikkeld om het probleem op te lossen?
   weinig -1-:-2-:-3-:-4- veel

2. In welke mate hebt u door het oplossen van dit PMP/PPP uw kennis met betrekking tot het oplossen van tandheelkundige problemen verruimd?
   weinig -1-:-2-:-3-:-4- veel

3. In welke mate ervaart u het oplossen van een PMP/PPP als prettig/plezierig?
   weinig -1-:-2-:-3-:-4- veel

4. In welke mate ervaart u het oplossen van het PMP/PPP als gemakkelijk?
   weinig -1-:-2-:-3-:-4- veel

5. In welke mate stond u duidelijk voor ogen hoe het PMP/PPP "in elkaar zat"?
   weinig -1-:-2-:-3-:-4- veel

6. In welke mate voelde u zich betrokken bij de PMP/PPP-patiënt en zijn problemen?
   weinig -1-:-2-:-3-:-4- veel

7. In welke mate kwam uw oplossing van het PMP/PPP op gestructureerde wijze tot stand?
   weinig -1-:-2-:-3-:-4- veel

8. In welke mate toetst volgens u een PMP/PPP probleemoplosvaardigheid?
   weinig -1-:-2-:-3-:-4- veel

9. In welke mate benaderde het oplosproces van het PMP/PPP volgens u het opstellen van een behandelingsplan in de realiteit?
   weinig -1-:-2-:-3-:-4- veel

CURRICULUM VITAE

De auteur van dit proefschrift werd geboren op 6 juli 1953 te Nijmegen. Na het behalen van de diploma's MULO-A en HAVO-A en de Akte Volledig Bevoegd Onderwijzer in respectievelijk 1969, 1971 en 1974, vervulde hij zijn militaire dienstplicht bij de Koninklijke Landmacht. In 1975 werd begonnen met de studie Pedagogische en Andragogische Wetenschappen aan de Katholieke Universiteit te Nijmegen. Het doctoraalexamen (afstudeerrichting "Onderwijskunde") werd afgelegd in 1980.
Sinds 1981 is hij als (tijdelijk) wetenschappelijk medewerker verbonden aan het Instituut Conserverende Tandheelkunde voor Volwassenen van de Katholieke Universiteit in Nijmegen. Zijn belangrijkste activiteiten liggen op het terrein van de evaluatie van het onderwijs.

STELLINGEN

behorende bij het proefschrift van

GERARD STRAETMANS:

"EVALUATIE IN HET TANDHEELKUNDIG ONDERWIJS"

NIJMEGEN, 10 MEI 1985

1. *Er is geen absolute norm voor de lengte van beoordelingsschalen. Veeleer is het aantal schaalpunten afhankelijk van het doel van de beoordeling (selectie of diagnose) en van de vaardigheid die beoordeeld wordt.*
(Dit proefschrift, deel I, hoofdstuk II)

2. *Prestatiecriteria zijn vaak dermate subjectief geformuleerd dat beoordelen veelal een interpreterende activiteit is.*
(Dit proefschrift, deel I, hoofdstuk III)

3. *De beoordelingskwaliteit van preklinische, tandheelkundige werkstukken is gebaat met het gebruik van een beoordelingsprotocol. Dit is een handleiding voor het beoordelen van practicumwerkstukken, bestaande uit: objectief geformuleerde prestatiecriteria, beoordelingsprocedures en scoringsregels.*
(Dit proefschrift, deel I, hoofdstuk III en VI)

4. *Er ontstaat een "kip-en-ei" probleem als de beoordelingskwaliteit beschreven wordt op basis van de overeenstemming met een expert-beoordeling.*
(Dit proefschrift, deel I, hoofdstuk VI)

5. *Papieren simulaties hebben het voordeel dat tandheelkunde studenten al in de beginfase van hun opleiding kunnen ervaren wat het belang is van de in het curriculum opgenomen vakken voor het oplossen van reële problemen.*
(Dit proefschrift, deel II, hoofdstuk I)

6. *Papieren patiënt problemen zijn, in tegenstelling tot patiënt management problemen, ongeschikt om te discrimineren tussen efficiënte en inefficiënte probleemoplossers.*
(Dit proefschrift, deel II, hoofdstuk II)

7. *Een "gecomputeriseerd patiënt management probleem" is meer dan een electronische versie van een patiënt management probleem.*
(Dit proefschrift, deel II, hoofdstuk IV)

8. *De afkortingen die gebruikt worden voor het aanduiden van nucleaire wapensystemen (ICBM = inter continental ballistic missile, SLBM = sea launched ballistic missile, GLCM = ground launched cruise missile, enz.) verhullen de dreiging die van dergelijke wapens uitgaat.*

9. *De explosieve groei van het aantal "off-the-road" voertuigen lijkt eerder veroorzaakt te worden door de wens om zich te onderscheiden dan door de kwaliteit van het Nederlandse wegennet.*

10. *De gretigheid waarmee sommige personen declareren doet vermoeden dat dit woord in de praktijk een ruimere betekenis heeft dan "het in rekening brengen van de gemaakte onkosten".*

11. *De nog altijd voortgaande zoutlozingen in de Rijn zijn een treffend voorbeeld van wat Montesquieu bedoelde met zijn uitspraak: "Souvent l'injustice n'est pas dans le jugement, elle est dans les délais."*

12. *Onderwijsvernieuwing, in de vorm van het invoeren van "interne differentiatie" als groeperingsvorm, is gedoemd te mislukken als het besluit hiertoe gebaseerd is op pragmatische (bijvoorbeeld minder onderwijzend personeel) in plaats van op onderwijskundige gronden.*

13. *Het discrimineren tussen nuchtere en aangeschoten voetgangers wordt ernstig bemoeilijkt door de sterke toename van het aantal honden en het daaruit voortvloeiend ongemak.*

14. *Vandaag is het grensverkeer aan onze oostelijke landsgrens aanzienlijk minder eenzijdig dan 45 jaar geleden.*